HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2024 March 07.

Published in final edited form as:

Nat Genet. 2023 October ; 55(10): 1640–1650. doi:10.1038/s41588-023-01497-6.

Genome-wide association meta-analysis identifies 17 loci associated with nonalcoholic fatty liver disease

Yanhua Chen[1,18], **Xiaomeng Du**[1,18], **Annapurna Kuppa**[1], **Mary F. Feitosa**[2], **Lawrence F. Bielak**[3], **Jeffrey R. O'Connell**[4], **Solomon K. Musani**[5], **Xiuqing Guo**[6], **Bratati Kahali**[1,7], **Vincent L. Chen**[1], **Albert V. Smith**[8], **Kathleen A. Ryan**[4], **Gudny Eirksdottir**[9], **Matthew A. Allison**[10], **Donald W. Bowden**[11], **Matthew J. Budoff**[12], **John Jeffrey Carr**[13], **Yii-Der I. Chen**[6], **Kent D. Taylor**[6], **Antonino Oliveri**[1], **Adolfo Correa**[5], **Breland F. Crudup**[5], **Sharon L. R. Kardia**[3], **Thomas H. Mosley Jr**[5], **Jill M. Norris**[14], **James G. Terry**[13], **Jerome I. Rotter**[6], **Lynne E. Wagenknecht**[15], **Brian D. Halligan**[1], **Kendra A. Young**[14], **John E. Hokanson**[14], **George R. Washko**[16], **Vilmundur Gudnason**[9,17], **Michael A. Province**[2], **Patricia A. Peyser**[3,19], **Nicholette D. Palmer**[11,19], **Elizabeth K. Speliotes**[1,19,✉]

[1]Department of Internal Medicine, Division of Gastroenterology and Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA.

[2]Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA.

[3]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA.

✉ **Correspondence and requests for materials** should be addressed to Elizabeth K. Speliotes. espeliot@med.umich.edu.

Author contributions

E.K.S. led the conceptualization, methodology development and funding of the project. P.A.P., N.D.P. and E.K.S provided supervision of the project. E.K.S., A.K. and N.D.P. led project management. Analysis were conducted by Y.C. (lead), X.D. (lead), B.K., M.F.F., L.F.B., K.A.R., S.K.M., K.A.Y., X.G., A.V.S., A.K., A.O., N.D.P. and B.F.C. Study resources were provided by N.D.P., D.W.B., L.E.W., J.R.O., S.K.M., K.D.T., S.L.R.K., T.H.M., A.C., J.I.R., V.G., J.M.N., M.A.P., P.A.P., J.E.H., G.R.W. and E.K.S. Data curation was performed by M.A.A., M.J.B., J.J.C., J.G.T., Y.-D.I.C., G.E., B.D.H. and E.K.S; Y.C., X.D., N.D.P., P.A.P. and E.K.S participated in central results interpretation. Paper draft preparation and editing was performed by E.K.S. (lead), Y.C., A.K., V.L.C., X.D., A.O. and N.D.P. Final review: all authors. All authors had access to the study data and reviewed and approved the final manuscript.

Competing interests

The Regents of the University of Michigan and E.K.S. have a pending patent on the use of systems and methods for analysis of samples associated with NAFLD and related conditions. V.L.C. received grant funding from KOWA and AstraZeneca. J.J.C. and Vanderbilt University Medical Center receive research funding from NIH, IBM Watson Health, GE Healthcare and Theratechnologies. G.R.W. is a cofounder and equity shareholder in Quantitative Imaging Solutions, a company that provides consulting services for image and data analytics. G.R.W.'s spouse works for Biogen. Grants or contracts from NIH, Department of Defense (DoD) and Boehringer Ingelheim made payments to G.R.W.'s institution. G.R.W. received consulting fees from Pulmonx, Vertex, Janssen Pharmaceuticals, Pieris Therapeutics and Intellia Therapeutics. G.R.W. also received payments from Pulmonx for participation on a Data Safety Monitoring Board or Advisory Board. The remaining authors declare no competing interests.

Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01497-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01497-6.

Peer review information *Nature Genetics* thanks Stefano Romeo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Code availability

Data analyses were performed using publicly available codes or software.

[4]Department of Endocrinology, Diabetes and Nutrition, University of Maryland – Baltimore, Baltimore, MD, USA.

[5]Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA.

[6]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor–UCLA Medical Center, Torrance, CA, USA.

[7]Centre for Brain Research, Indian Institute of Science, Bangalore, India.

[8]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA.

[9]Icelandic Heart Association, Kopavogur, Iceland.

[10]Department of Family Medicine, University of California San Diego, San Diego, CA, USA.

[11]Department of Biochemistry, Wake Forest University School of Medicine, Winston-Salem, NC, USA.

[12]Department of Internal Medicine, Lundquist Institute at Harbor–UCLA, Torrance, CA, USA.

[13]Department of Radiology, Vanderbilt University School of Medicine, Nashville, TN, USA.

[14]Department of Epidemiology, Colorado School of Public Health, Aurora, CO, USA.

[15]Division of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, NC, USA.

[16]Department of Medicine, Division of Pulmonary and Critical Care, Brigham and Women's Hospital, Boston, MA, USA.

[17]Department of Medicine, University of Iceland, Reykjavik, Iceland.

[18]These authors contributed equally: Yanhua Chen, Xiaomeng Du.

[19]These authors jointly supervised this work: Patricia A. Peyser, Nicholette D. Palmer, Elizabeth K. Speliotes.

Abstract

Nonalcoholic fatty liver disease (NAFLD) is common and partially heritable and has no effective treatments. We carried out a genome-wide association study (GWAS) meta-analysis of imaging (n = 66,814) and diagnostic code (3,584 cases versus 621,081 controls) measured NAFLD across diverse ancestries. We identified NAFLD-associated variants at torsin family 1 member B (*TOR1B*), fat mass and obesity associated (*FTO*), cordon-bleu WH2 repeat protein like 1 (*COBLL1*)/growth factor receptor-bound protein 14 (*GRB14*), insulin receptor (*INSR*), sterol regulatory element-binding transcription factor 1 (*SREBF1*) and patatin-like phospholipase domain-containing protein 2 (*PNPLA2*), as well as validated NAFLD-associated variants at patatin-like phospholipase domain-containing protein 3 (*PNPLA3*), transmembrane 6 superfamily 2 (*TM6SF2*), apolipoprotein E (*APOE*), glucokinase regulator (*GCKR*), tribbles homolog 1 (*TRIB1*), glycerol-3-phosphate acyltransferase (*GPAM*), mitochondrial amidoxime-reducing component 1 (*MARC1*), microsomal triglyceride transfer protein large subunit (*MTTP*), alcohol dehydrogenase 1B (*ADH1B*), transmembrane channel like 4 (TMC4)/membrane-bound

O-acyltransferase domain containing 7 (*MBOAT7*) and receptor-type tyrosine-protein phosphatase δ (*PTPRD*). Implicated genes highlight mitochondrial, cholesterol and de novo lipogenesis as causally contributing to NAFLD predisposition. Phenome-wide association study (PheWAS) analyses suggest at least seven subtypes of NAFLD. Individuals in the top 10% and 1% of genetic risk have a 2.5-fold to 6-fold increased risk of NAFLD, cirrhosis and hepatocellular carcinoma. These genetic variants identify subtypes of NAFLD, improve estimates of disease risk and can guide the development of targeted therapeutics.

With rising obesity rates, the prevalence of nonalcoholic fatty liver disease (NAFLD) has increased to epidemic proportions. NAFLD is caused by the deposition of excess fat in the liver (not due to alcohol) and can lead to advanced liver diseases, including inflammation, fibrosis/cirrhosis (scarring) and hepatocellular carcinoma (HCC; liver cancer)[1]. NAFLD is associated with metabolic diseases including dy slipidemia, hypertension, cardiovascular disease and diabetes, although causal relationships have not been established[2–8]. More than 90% of severely obese individuals suffer from advanced NAFLD, which is associated with decreased lifespan[9]. The disease imposes an annual direct medical cost of approximately $103 billion in the United States and will soon become the leading indication for liver transplantation[10]. Causes of NAFLD are poorly understood, and there are presently no effective treatments, making this a large unmet medical need.

We and others have shown that NAFLD is partially heritable and have identified variants associated with disease[8,11–20]. These variants explain only about 20% of the heritability, suggesting that additional genetic risk variants remain to be identified. NAFLD steatosis can be measured using computed tomography (CT) or magnetic resonance imaging (MRI), which have an r^2 with histological steatosis of 0.78 and 0.98, respectively[21,22]. Through the use of electronic health records, cases and controls can now be identified by International Classification of Diseases (ICD) codes and using natural language processing (NLP) of imaging and pathology reports. Furthermore, availability of liver imaging in large population-based cohorts and biobanks now allows direct assessment of these populations for the study of NAFLD and its associated comorbidities.

Results

GOLDPlus meta-analysis

We carried out a multi-ancestry meta-analysis of CT-measured liver fat (Genetics of Obesity-related Liver Disease (GOLD)) with UK Biobank (UKBB) MRI liver proton density fat fraction (PDFF), UKBB NAFLD, Electronic Medical Record and Genomics (eMERGE) NAFLD and FinnGen NAFLD (Extended Data Fig. 1). In all analyses, the top associated variants were at patatin-like phospholipase domain-containing protein 3 (*PNPLA3*), verifying congruency across phenotypes. We identified 17 independent genome-wide significant variants ($P < 5.00 \times 10^{-8}$; Table 1 and Supplementary Fig. 1). We prioritized genes for annotation if the index variant was an exonic variant, was in high LD (linkage disequilibrium, $r^2 > 0.85$) with an exonic variant, and/or was an expression quantitative trait loci (eQTL) for the gene expressed in liver. Genes that were within 1 Mb of the index variant and predominantly expressed in liver, prioritized by data-driven

expression prioritization integration for complex traits (DEPICT) analysis, and/or nearest to the index variant were also prioritized for annotation[23]. One region contained possibly two independent loci within close proximity—alcohol dehydrogenase 1B (*ADH1B*)-rs1229984, which is within 500 kb of microsomal triglyceride transfer protein large subunit (*MTTP*)-rs7661964. To determine whether these two signals were independent, we carried out conditional analyses in the UKBB multi-ancestry dataset. *ADH1B*-rs1229984 had *P* values of 5.09×10^{-6} and 1.03×10^{-5} before and after conditioning on *MTTP*-rs7661964, respectively. *MTTP*-rs7661964 had *P* values of 2.01×10^{-7} and 4.09×10^{-7} before and after conditioning on *ADH1B*-rs1229984, respectively. We defined novel variants as those more than 1 Mb from published NAFLD or hepatic steatosis genome-wide association study (GWAS) genome-wide significant variants ($P < 5.00 \times 10^{-8}$) at the time of manuscript submission. We identified novel associations in or near torsin family 1 member B (*TOR1B*), fat mass and obesity associated (*FTO*), cordon-bleu WH2 repeat protein like 1 (*COBLL1*)/growth factor receptor-bound protein 14 (*GRB14*), insulin receptor (*INSR*), sterol regulatory element-binding transcription factor 1 (*SREBF1*), and patatin-like phospholipase domain-containing protein 2 (*PNPLA2*; Table 1 and Supplementary Fig. 1). We confirmed previously identified NAFLD associations in or near *PNPLA3*, transmembrane 6 superfamily 2 (*TM6SF2*), apolipoprotein E (*APOE*), glucokinase regulator (*GCKR*), tribbles homolog 1 (*TRIB1*), glycerol-3-phosphate acyltransferase (*GPAM*), mitochondrial amidoxime-reducing component 1 (*MARC1*), *MTTP*, *ADH1B*, transmembrane channel like 4 (*TMC4*)/membrane-bound O-acyltransferase domain containing 7 (*MBOAT7*) and receptor-type tyrosine-protein phosphatase δ (*PTPRD*)[8,11–20]. One variant *LOC157273*/protein phosphatase 1 regulatory subunit 3B (*PPP1R3B*)-rs4841132 ($P = 4.21 \times 10^{-13}$; $P_{het} = 7.44 \times 10^{-19}$) was removed from downstream analysis due to phenotype heterogeneity (Methods). Rs4841132 is known to promote liver damage by increasing glycogen, which is a distinct pathology from NAFLD[24]. In a separate European (EUR) ancestry-only meta-analysis (Extended Data Fig. 2), only the *PTPRD* locus was not genome-wide significant (Supplementary Tables 1 and 2), likely due to reduced power as we do not see heterogeneity of effect across ancestries.

Effects of variants by study, ancestry, sex and alcohol use

We assessed heterogeneity of effect for NAFLD-associated variants in GOLDPlus. After Bonferroni correction, only *TM6SF2*-rs58542926 and *APOE*-rs429358 showed statistically significant heterogeneity of effect. However, direction of effect across studies was congruent. For completeness, we show the effects of the loci overall and stratified by cohort (Table 1 and Supplementary Table 3, respectively).

We next assessed the effects of NAFLD-associated variants across ancestries (EUR, n = 15,880; African (AFR), n = 5,607; Hispanic (HIS), n = 1,674 and Chinese (CHN), n = 360; Fig. 1 and Extended Data Fig. 3) and sex (males, n = 11,006; females, n = 12,515; Extended Data Fig. 4). For these analyses, we used the GOLD Consortium data, where we had the highest quality measures of hepatic steatosis in population-based cohorts (Supplementary Tables 4–10). *PNPLA3* (β = 0.24 EUR, β = 0.27 AFR, β = 0.24 HIS, β = 0.17 CHN, $P_{het} = 5.69 \times 10^{-6}$) exhibited significant heterogeneity of effect across ancestries. However, a limited sample size in the CHN-ancestry cohort likely caused an unstable estimate of

beta, influencing the estimate of heterogeneity. After the removal of the CHN cohort from the meta-analysis, the heterogeneity *P* value was not significant after Bonferroni correction (*PNPLA3*, $P_{het} = 0.69$). No other loci showed significant heterogeneity of effect by ancestry or sex.

We found >10% absolute difference in effect allele frequencies (EAFs) for index variants in *PNPLA3*, *GCKR*, *TRIB1*, *GPAM*, *MARC1*, *ADH1B*, *MTTP*, *FTO*, *INSR*, *TMC4/MBOAT7*, *SREBF1* and *PTPRD* across ancestries (Fig. 1 and Extended Data Fig. 3). The starkest contrast in allele frequencies across ancestries existed in *ADH1B*. In the CHN-ancestry cohort, *ADH1B* (rs1229984-C) had an EAF of 0.26, while it had >91% EAF in EUR-, AFR-, and HIS-ancestry cohorts. The variance explained across ancestries paralleled allele frequencies more than effect sizes, which were similar across ancestries (Supplementary Table 10). The highest variances explained were 2.79% in the HIS cohort for *PNPLA3*, 2.42% in the CHN cohort for *GCKR* and 2.04% in the EUR cohort for *PNPLA3* (Supplementary Table 10). Taken together, these findings suggest EAF, more than effect size, accounts for differences in genetic disease burden across ancestries.

To assess the effects of alcohol use, we used our largest population-based cohort, UKBB MRI-PDFF, to perform a GWAS analysis stratified by alcohol use (Supplementary Table 11). After Bonferroni correction, only *ADH1B* exhibited significant heterogeneity of effect ($P_{het} = 6.16 \times 10^{-4}$) between heavy (≥14 drinks per week for males or ≥7 drinks a week for females; $n = 21,356$) and light (≤1 drinks per week for males and females; $n = 9,871$) drinkers (Supplementary Table 12). *ADH1B* had a significantly greater effect ($\beta = 0.20$) in heavy drinkers compared to light drinkers ($\beta = 0.03$).

Tissue, gene set and pathway analyses

To further understand the biology underlying NAFLD associations, we used DEPICT to identify enriched tissues and cell types (false discovery rate (FDR) $P < 0.05$)[23]. Input included the 17 NAFLD-associated variants. Liver and adipose tissue were the most enriched tissues (Extended Data Fig. 5). Epithelial cells (hepatocytes) were the most enriched cell type (Extended Data Fig. 5). Using mSigDB, we computed significant gene functional overlaps[25]. We found enrichment (FDR $P < 0.01$) in the following biological functions: lipid homeostasis, lipid metabolic processes, monocarboxylic acid metabolic processes, alcohol metabolic processes, lipid biosynthesis, regulation of cholesterol biosynthesis and steroid biosynthesis.

Association of NAFLD variants with other phenotypes

We used publicly available GWAS data to perform a phenome-wide association study (PheWAS) of NAFLD-risk-increasing alleles with ICD-based diseases, alcohol intake, cardiovascular and body composition measures, and lipid, metabolic and liver function tests (Fig. 2). The NAFLD-risk-increasing allele of the variants broadly separated into the following two groups: one showing significant associations with increased serum low-density lipoprotein cholesterol (LDL) and increased alanine aminotransferase (ALT; *TRIB1*, *GCKR*, *COBLL1/GRB14*, *INSR*, *PNPLA2*, *SREBF1*, *MTTP*, *GPAM*, *MARC1*, *TMC4/MBOAT7*, *TOR1B* and *ADH1B* associations) and one exhibiting decreased associations

with LDL and increased associations with ALT (*FTO*, *PTPRD*, *PNPLA3*, *TM6SF2* and *APOE*). Further separations showed that variants at *TRIB1*, *GCKR*, *COBLL1/GRB14*, *INSR*, *PNPLA2* and *SREBF1* were distinguished from *GPAM*, *MARC1*, *TMC4/MBOAT7*, *TOR1B* and *ADH1B* by being associated with high triglycerides (TG) and low high-density lipoprotein cholesterol (HDL). NAFLD-associated variants at *TRIB1* and *GCKR* were distinguished from *COBLL1/GRB14*, *INSR*, *PNPLA2*, *SREBF1* and *MTTP* by being associated with low risk of cholelithiasis and cholecystitis. *GCKR* had a particularly strong association with lower insulin-like growth factor 1 (IGF-1) and sex hormone-binding globulin (SHBG) levels. NAFLD-increasing associations at *FTO* were associated with increased TG, whereas those at *PTPRD*, *PNPLA3*, *TM6SF2* and *APOE* were associated with decreased TG. *FTO* clustered alone and differed from other loci in having very strong association with increased body mass index (BMI). Likewise, *APOE* clustered alone and differed from *PNPLA3* and *TM6SF2* associations in having an increased association with body composition measures and decreased association with familial Alzheimer's disease. We also looked at the effect of PheWAS subgroupings on diseases/traits in UKBB and depicted them as forest plots showing associations between subgroups and human diseases/traits (Fig. 3). We found similar results as mentioned above, where the variants grouped in the glucose, insulin, absorb and TG divert categories had increased LDL and ALT, while those grouped as low lipid burn, low output and high input had increased ALT but decreased LDL. Variants grouped as glucose and insulin differed from those in TG divert by being associated with high TG and low HDL. Those grouped in categories of low output and high input were associated with decreased TG. Variants grouped in the high input category differed from the low output by having an increased association with BMI. A schematic providing biological context for the PheWAS subgroupings is presented in Fig. 4.

Association of NAFLD polygenic risk score (PRS) with other human phenotypes

To assess the cumulative effects of NAFLD-risk-increasing variants on disease, we constructed a PRS based on the GOLD-weighted (multi-ancestry) NAFLD-associated single variants (n = 17) and performed a PheWAS using ICD-9 and ICD-10 diagnoses available in an independent cohort, that is, Michigan Genomics Initiative (MGI; Fig. 5; characteristics in Supplementary Table 13). The PRS strongly associated with digestive phenotypes, that is, other chronic nonalcoholic liver diseases (odds ratio (OR) = 1.38 (95% confidence interval (CI) = 1.33–1.43)) as the most significant, chronic liver disease and cirrhosis (OR = 1.37 (95% CI = 1.31–1.42)), cirrhosis without mention of alcohol (OR = 1.48 (95% CI = 1.39–1.59)), liver abscess and sequelae of chronic liver disease (OR = 1.52 (95% CI = 1.39–1.66)), other disorders of liver (OR = 1.24 (95% CI = 1.18–1.29)), abnormal results of liver function (OR = 1.24 (95% CI = 1.17–1.31)), portal hypertension (OR = 1.51 (95% CI = 1.36–1.69)), esophageal bleeding (OR = 1.46 (95% CI = 1.32–1.62)), abnormal serum enzyme levels (OR = 1.16 (95% CI = 1.11–1.21)), nonmalignant ascites (OR = 1.24 (95% CI = 1.15–1.34)) and liver transplant (OR = 1.45 (95% CI = 1.27–1.67)) as the least significant digestive phenotype. We also found significant associations with malignant liver neoplasm (OR = 1.60 (95% CI = 1.39–1.83)), cancer of liver and intrahepatic bile duct (OR = 1.35 (95% CI = 1.21–1.50)), type 2 diabetes (OR = 1.07 (95% CI = 1.04–1.09)) and alcoholic liver damage (OR = 1.65 (95% CI = 1.45–1.89)).

Effects on liver outcomes by PRS percentile

We then examined the effect of the PRS on NAFLD, cirrhosis and HCC by PRS percentiles. Higher NAFLD PRS was strongly associated with an increased OR for NAFLD in MGI (Fig. 6a). Compared to those in the PRS bottom decile, individuals in the top 10%, 5% and 1% had an OR = 2.79 (95% CI = 2.36–3.29), 3.46 (95% CI = 2.88–4.15), and 4.77 (95% CI = 3.62–6.27), respectively, for NAFLD. Higher NAFLD PRS was also associated with increased odds of both cirrhosis (top 10% OR = 2.51 (95% CI = 1.98–3.17), top 5% OR = 3.43 (95% CI = 2.66–4.41) and top 1% OR = 5.14 (95% CI = 3.59–7.36)) and HCC (top 10% OR = 2.89 (95% CI = 1.76–4.76), top 5% OR = 4.25 (95% CI = 2.53–7.16) and top 1% OR = 5.80 (95% CI = 2.83–11.92); Fig. 6b,c) in MGI.

Mendelian randomization (MR)

To determine whether NAFLD causally influences liver and metabolic diseases and traits, we performed two-sample MR using variant-NAFLD effect estimates from GOLD as the exposure and related publicly available and UKBB GWAS as the outcome. NAFLD-associated variants with an *F*-statistic >10 were used as a combined instrumental variable for steatosis ($n = 11$; Supplementary Table 14; combined *F* statistic = 45.4)[26]. Using the GOLD effects as the exposure, we found NAFLD increased the risk of liver fibrosis and cirrhosis (ICD K74, OR = 1.002, 95% CI = 1.001–1.003; MR-Egger, $P = 1.88 \times 10^{-3}$; OR = 1.001, 95% CI = 1.001–1.002; inverse-variance weighted (IVW), $P = 8.65 \times 10^{-5}$) and esophageal varices (ICD I85; OR = 1.003, 95% CI = 1.002–1.005, MR-Egger $P = 9.36 \times 10^{-4}$; OR = 1.002, 95% CI = 1.001–1.003, IVW $P = 3.51 \times 10^{-4}$; Extended Data Fig. 6). The MR-Egger heterogeneity *P* values were not significant for fibrosis ($P_{het} = 0.32$) but were significant for esophageal varices ($P_{het} = 0.02$). The MR-Egger pleiotropy *P* values were not significant for fibrosis ($P = 0.08$) but were significant for esophageal varices ($P = 0.03$), indicating horizontal pleiotropy may be driving the results of the esophageal varices MR. Sensitivity analyses are shown in Extended Data Fig. 6c,d.

We then assessed the causal effects of metabolic disorders, body composition measures and advanced liver disease on NAFLD. We used the GOLD all-ancestry meta-analysis as the outcome and independent genome-wide significant variants ($P < 5.00 \times 10^{-8}$) from previously published GWAS (Supplementary Table 15) as the exposure. Increased BMI (OR = 1.29, 95% CI = 1.05–1.59, MR-Egger $P = 0.02$; OR = 1.203, 95% CI = 1.12–1.29, IVW $P = 1.02 \times 10^{-7}$) and waist circumference (OR = 1.36, 95% CI = 1.02–1.82, MR-Egger $P = 3.6 \times 10^{-2}$; OR = 1.18, 95% CI = 1.08–1.29, IVW $P = 3.71 \times 10^{-4}$) increased risk of NAFLD (Extended Data Fig. 7). The MR-Egger heterogeneity and pleiotropy *P* values were not significant for BMI and waist circumference. The sensitivity analyses are shown in Extended Data Figure 7c,d.

Discussion

The present study represents the largest GWAS meta-analysis of CT-measured, MRI-measured and diagnostic-code-assessed NAFLD to date. We identified 17 loci that include new genes associated with NAFLD. The effects of these variants on NAFLD were congruent across study, ancestry and sex. However, some of the associated variants have EAF

differences across ancestries, which were consistent with differences in the population burden of NAFLD. One variant, *ADH1B*-rs1229984, had a substantially varied effect when stratified by alcohol consumption. Tissue and pathway enrichment analyses identified liver, lipid, cholesterol, steroid, alcohol and monocarboxylic acid processes as being enriched. PheWAS analysis highlighted at least seven subtypes/clusters of NAFLD-associated variants and implicated genes that have a role in mitochondrial, very low-density lipoprotein cholesterol (VLDL), cholesterol and de novo lipogenesis processes. A PRS of the NAFLD-associated genetic variants can identify people with an elevated risk of NAFLD, cirrhosis and HCC.

Our approach combining imaging, ICD-based and NLP-based diagnosis of NAFLD offers substantial advantages over traditional histology- or single modality-based GWAS. The use of ICD-based diagnosis could underestimate the disease; however, this causes the statistics to move toward the null so that any positive associations are still valid and interpretable. These nonhistology measures are less expensive, less invasive and more ethically applicable to asymptomatic individuals than liver biopsy. The inclusion of nonhistology-measured NAFLD increases power and decreases ascertainment bias. Furthermore, by assessing the heterogeneous effects of variants across multiple modalities, we were able to identify a variant associated with other types of liver disease, such as glycogen storage disease, that can be misdiagnosed as NAFLD and remove it from analysis. Thus, to reduce possible misdiagnosis, decrease ascertainment bias and increase power, we propose using multiple methods for measuring NAFLD when available.

One method used to assess fatty liver is abdominal MRI. To increase the power of our analysis, we used a convolutional neural network (CNN) to train an algorithm to measure MRI-PDFF in all UKBB participants with abdominal MRI (Supplementary Table 16). We then validated and tested the computed values in overlapping samples with UKBB MRI-PDFFs measured by expert radiologists. GWAS with the CNN-derived MRI-PDFFs identified variants near *PNPLA3*, *TM6SF2*, *APOE* and *GCKR* as the most strongly associated, verifying the sensitivity and specificity of the measure (Supplementary Table 17). GWAS with just 4,616 samples where the MRI-PDFF value was known gave an association *P* value for *PNPLA3* of 2.61×10^{-24}, whereas GWAS of 43,293 samples with CNN-derived MRI-PDFF values decreased the *P* value to 2.18×10^{-132}. Combining machine-learning methods with expert MRI or other imaging measures greatly increases the efficiency with which we can analyze large numbers of imaging studies present in biobanks and medical record archives. Thus, this combination of tools can facilitate larger sample sets to study the genetic epidemiology of many human diseases and traits.

The power of our meta-analysis allowed us to identify several genome-wide significant variants associated with hepatic steatosis and NAFLD that were new at the time of manuscript submission, including *TOR1B*, *FTO*, *COBLL1/GRB14*, *INSR*, *SREBF1* and *PNPLA2*. Our group and others have shown *TOR1B* to be associated with elevated serum liver enzymes[27] and cirrhosis[28], but here we additionally show a strong association with direct measures of hepatic steatosis and NAFLD. Previously implicated variants that were confirmed by our analyses include *PNPLA3*, *TM6SF2*, *APOE*, *GCKR*, *TRIB1*, *GPAM*, *MARC1*, *MTTP*, *ADH1B*, *TMC4/MBOAT7* and *PTPRD*[8,11–20]. Since the time

of manuscript submission, recent publications[29,30] have also reported NAFLD-associated variants in or near *TOR1B*, *FTO*, *COBLL1* and *PNPLA2*, reinforcing our studies.

We identified seven distinct clusters among the NAFLD variants and their associations with related phenotypes and cellular localization of the resulting protein product. NAFLD-promoting alleles in *TRIB1* and *GCKR* were associated with increased TG levels in our PheWAS. *GCKR* is an established variant associated with hepatic steatosis[12]. *TRIB1* has been previously associated with lipid levels, myocardial infarction and liver function tests[27,31,32]. Both TRIB1 and GCKR are thought to use glucose to promote de novo lipogenesis in liver, which may explain the identical associations with hepatic steatosis and related phenotypes[33,34]. Mice lacking TRIB1 have increased fatty acid synthesis and incorporation of diacyl to triacyl glycerol[35]. GCKR negatively regulates glucokinase; in the absence of GCKR, glucokinase promotes many carbohydrate-responsive processes, including promoting de novo lipogenesis[36]. In our PheWAS, we observed that NAFLD-promoting variants at *GCKR* and *TRIB1* were associated with increased cholesterol and risk for myocardial infarction. These are the changes that would be predicted to occur if there was a loss of function at these genes that promoted de novo lipogenesis and cholesterol synthesis. The NAFLD-promoting variant at *GCKR*, however, was strongly associated with lower glucose and lower prevalence of diabetes, whereas the NAFLD-promoting variant at *TRIB1* was neutral in this regard. This suggests that the loss of these genes does not result in molecular changes that are completely convergent.

Like *TRIB1* and *GCKR*, NAFLD-increasing variants in *COBLL1/GRB14*, *INSR*, *PNPLA2*, *SREBF1* and *MTTP* were associated with increased TG, but unlike *TRIB1* and *GCKR* were associated with increased glucose and glycated hemoglobin. Heterozygous mutations in *PNPLA2* have been shown to cause a Mendelian disease, that is, neutral lipid storage disease (NLD). NLD is characterized by severe accumulation of cytoplasmic lipid droplets and associated with abnormal lipid metabolism in multiple tissues with muscle weakness, insulin resistance and diabetes and hepatic steatosis[37]. *SREBF1* is a transcription factor activated by insulin that regulates hepatic de novo lipogenesis[38]. Experimental evidence has shown that increased expression of *SREBF1* contributes to NAFLD via de novo lipogenesis[39]. *GRB14* is a negative regulator of insulin signaling. Downregulation of *Grb14* in mice decreased blood glucose and improved liver steatosis[40,41]. The metabolic effects of insulin are mediated through *INSR*. Defects in *INSR* impair the biological response to insulin and lead to insulin resistance[42]. Insulin resistance has been shown to promote hepatic steatosis, perhaps by causing the increased release of free fatty acids from adipose tissue that can go to the liver to increase fatty liver[42]. *MTTP* is a well-known gene whose product transfers phospholipids and triacylglycerols to nascent apoB for the assembly of lipoproteins, has previously been associated with NAFLD, and whose absence causes the Mendelian disease abetalipo-proteinemia[43,44]. *MTTP*, unlike others in this set, did not associate with decreased HDL and BMI, reflecting different biology. This is likely due to effects in the intestine, where it helps in absorption and packaging of lipids; the other members of this group were related to insulin biology.

In contrast to the abovementioned genes, NAFLD-promoting variants in *GPAM*, *MARC1*, *TMC4*/*MBOAT7*, *TOR1B* and *ADH1B* associated with low TG but high LDL, HDL,

ischemic heart disease and hypertension. The functions of steatosis-promoting alleles (*GPAM*, *ADH1B* and *TOR1B*) include TG synthesis and alcohol/retinol metabolism. GPAT1 catalyzes the first step in glycerolipid biosynthesis and promotes TG synthesis. Both *GPAM* and *ADH1B* are highly expressed in adipose and liver tissue[45]. ADH1B is a member of the alcohol dehydrogenase family whose members metabolize many substrates, including ethanol, retinol, aliphatic alcohols, hydroxysteroids and lipid peroxidation products, and has previously been shown to be associated with NAFLD[19,46]. Notably, the variant identified in *ADH1B* shows positive selection in East Asians[47] and is associated with alcohol use[48]. People without this allele are not able to metabolize alcohol well, have adverse reactions to alcohol and are less likely to become alcohol-dependent. Other substrates that ADH1B could affect to promote NAFLD and advanced liver disease include retinol, which when present in high amounts promotes cirrhosis[49]. Much less is known about the function of *TOR1B*. However, conditional deletions of torsins in mouse hepatocytes resulted in profound lipid dysregulation, which included increased intracellular lipids, decreased LDL/TG secretion, and decreased circulating lipids[50]. The *MARC1* fatty liver-promoting allele has been shown to increase NAFLD and alcoholic cirrhosis, but its physiological function on NAFLD remains unclear[13,51]. Liver-specific knockout of *Lpiat1*, the mouse homolog of *MBOAT7*, altered the fatty acid content of phosphoinositols and activated SREBP-1c to increase de novo lipogenesis and liver inflammation[52]. Induction of SREBP-1c alone, however, would not explain why individuals carrying the *MBOAT7* NAFLD-promoting allele have decreased TG and increased LDL and HDL, suggesting that other molecular details of its action remain to be elucidated. Overall, our results suggest that these variants (*GPAM*, *ADH1B*, *TOR1B*, *MARC1* and *MBOAT7*) may function in similar ways to divert/reduce TG from serum and increase TG levels, fatty acid content of phospholipids, and other substrates in the liver, leading to NAFLD, although their exact mechanisms remain largely unknown.

Unlike the abovementioned genes, the NAFLD-increasing variants in *FTO* associated with increased TG, although modestly, and decreased LDL. In the PheWAS, *FTO* is most strongly associated with increased BMI and overall obesity measures and increased glucose and diabetes. This pattern of associations was unique, resulting in its clustering by itself. Variants at *FTO* likely act via *IRX3/5* to decrease adipocyte browning, reduce lipid burning and increase BMI; whether the effect on fatty liver is direct in liver versus indirect by affecting adipose tissue remains to be determined[53].

PTPRD, PNPLA3 and *TM6SF2* reduce LDL but are also associated with decreased, as opposed to increased, BMI. PTPRD has been shown to have a role in hepatic lipid accumulation through exacerbation of the dephosphorylation of tyrosine 705 of the signal transducer and activator of transcription 3 protein (pSTAT3) in hepatocytes[17]. Mutations in *TM6SF2* and *PNPLA3* are thought to reduce serum lipids by decreasing their release from hepatocytes. TM6SF2 has been proposed to affect the lipid loading of VLDL, whereas PNPLA3 I148M has been proposed to affect the release of lipids from lipid droplets to cause lipid accumulation in liver[54,55]. Overall, this group of proteins results in decreased lipoprotein cargo output to cause NAFLD.

Finally, APOE binds to multiple liver receptors facilitating the uptake of HDL and thereby increasing lipoprotein cargo input to liver[56]. It also helps metabolize retinols and thus

may serve dual roles in promoting steatosis and progression to advanced liver diseases[18]. The fatty liver-promoting allele at *APOE*, unlike others, associates with reduced risk of Alzheimer's disease and increased systemic C-reactive protein. Interestingly, the other variant at *APOE* that promotes Alzheimer's disease (rs7412) does not have a strong association with fatty liver, thus demonstrating allele specificity[18].

PRS analysis of these PheWAS subtypes showed that they have different effects on cardiometabolic and gastrointestinal (GI) diseases. This helps explain different patterns of disease associations and helps link these to genes and pathways that may be targeted in the future for precision therapeutics.

Some of the genes that have been previously suggested to have a role in NAFLD were not found to be significantly associated with hepatic steatosis measures in GOLDPlus. This includes *HSD17B13* (rs6834314; $P = 0.64$), which may have a stronger effect on NASH/fibrosis. *ERLIN1/CHUK/CWF19L1* (rs17729876; $P = 1.15 \times 10^{-5}$) and *LYPLAL1* (rs12137855; $P = 0.002$) had weak effects on promoting NAFLD. *LYPLAL1* may affect body fat distribution and indirectly affect NAFLD, helping to explain its particularly weak effect[57,58].

In addition to identifying new variants associated with NAFLD, our study examined the combined effect of single variants using MR, pathway analysis and PRS. MR analysis suggests that obesity is causally related to the development of NAFLD, but not the reverse. However, MR showed hepatic steatosis is causally related to fibrosis/cirrhosis. Considering these effects, hepatic steatosis should be targeted for intervention before the development of advanced liver disease, and efforts to combat obesity must be continued. As expected, tissues, cell types and pathway analyses highlight liver and then adipose as key tissues. Pathway enrichment analyses identified lipid, cholesterol, steroid, alcohol and monocarboxylic acid processes as being enriched for association with NAFLD. These results highlight the complex biological etiologies of NAFLD. Finally, our results demonstrate that, taken together, these 17 genetic variants can identify patterns of disease expected in individuals that carry these variants in an independent cohort. Indeed, the PRS was able to identify individuals at high risk (OR > 2) of NAFLD, cirrhosis and HCC in the top 5% of the PRS so that they can be targeted for interventions to prevent the development of advanced liver disease.

Our study is limited by the fact that it does not use histology to define all components of NAFLD. Additionally, most of our study group was EUR ancestry, so our predictions are likely to be most accurate in this population. Also, for publicly available datasets (FinnGen and eMERGE), we did not adjust for confounders as we lacked access to the primary data.

Our study has many strengths and represents the largest genetic study of hepatic steatosis to date, allowing us to identify new variants. Here we created and validated new machine-learning methods to predict MRI-PDFF from abdominal MRI, which can be used to facilitate future studies incorporating imaging analysis for NAFLD and other endpoints. We used multiple methods to diagnose NAFLD, allowing us to identify variants that associate with misdiagnosed NAFLD in the population, namely *PPP1R3B*. We assessed our variants

of interest across cohorts, ancestry, sex and alcohol consumption and found congruent effects. EAF determined disease burden across ancestries. We assessed the effects of variants across many human diseases and traits and used MR to determine which traits promoted NAFLD and which were promoted by NAFLD. We created a PRS for NAFLD that predicts hepatic steatosis and cirrhosis. Finally, we identified NAFLD disease subtypes that can be linked to specific biology and targeted for future NAFLD precision therapy.

In conclusion, we identified 17 loci associated with hepatic steatosis/NAFLD. These single-variant analyses identify genes that have a role in hepatic lipid processes, and the variants have diverse effects on human traits, suggesting that targeted interventions may be needed to effectively treat individuals with various disease subtypes. Genetic testing with these variants or combinations may help identify individuals at increased risk of developing advanced liver disease more effectively than clinical measures alone. MR shows that hepatic steatosis is causally related to the development of advanced liver disease and should be treated to prevent disease progression.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01497-6.

Methods

Ethical approvals and study design

Protocols were approved by the institutional review boards (IRBs) at the institutions where participants were recruited. All included participants provided written informed consent. All research in this study was approved by the IRB of the University of Michigan (Ann Arbor, MI). UKBB protocols were approved by the National Research Ethics Service Committee, and all participants provided written informed consent. Analyses in this project were conducted under UKBB Resource Project 18120. IRB approval was not required to use eMERGE and FinnGen data as they are publicly available. Analyses were carried out in cohorts from the GOLD Consortium, UKBB, FinnGen, eMERGE consortium and MGI (Extended Data Fig. 1).

GOLD Consortium

The multi-ethnic GOLD Consortium includes the following nine multi-ethnic cohorts with CT-measured steatosis (n = 23,521): Age, Gene/Environment Susceptibility-Reykjavik Study (AGES)[75], COPDGene[76], Family Heart Study (FamHS)[77], Framingham Heart Study (FHS)[78], Genetic Epidemiology Network of Arteriopathy (GENOA)[79], Insulin Resistance Atherosclerosis Family Study (IRASFS)[80], Jackson Heart Study (JHS)[81], Multiethnic Study of Atherosclerosis (MESA)[82] and Old Order Amish (OOA)[83]. The normalization of liver attenuation and control for scan penetrance protocols, cohort characteristics and details of the genotyping methods, quality control and sample exclusion criteria are provided in

Supplementary Tables 4–8. GOLD genotypes were called in the individual datasets using Illumina Bead-Studio, GenomeStudio, BRLMM or Birdseed.

UKBB

The UKBB cohort was previously described[84]. Participants in the NAFLD analyses were included regardless of ethnicity and excluded if they or their relatives had abdominal MRI images. NAFLD cases were identified by ICD-9 571.8 or ICD-10 K76.0 codes. The UKBB NAFLD dataset included 1,827 NAFLD cases and 436,262 controls (Supplementary Table 18). A second UKBB NAFLD EUR-only dataset was assembled as stated above and included 1,706 cases and 412,151 controls (Supplementary Table 18). For quality control of UKBB genotypic data, we used EasyQC 9.2.

CNN model for UKBB liver MRI imaging

We applied a CNN model to determine liver PDFF from MRI in UKBB. UKBB uses the following two imaging protocols: gradient echo (GRE; $n = 10,093$) and iterative decomposition of water and fat with echo assymtery and least-squares estimation (IDEAL; $n = 35,779$), which includes $n = 1,491$ individuals who had undergone both protocols. To determine the MRI-PDFF for all participants, we applied a standard 2D U-Net to segment the GRE and IDEAL liver data[85]. We used ITK-SNAP (version 3.8) software to manually annotate the liver in 98 randomly chosen images from the GRE protocol[86]. Next, we split the segmented GRE images into training ($n = 64$), validation ($n = 16$) and test ($n = 18$) sets. The result showed that liver segmentation achieved Dice scores over 94%. Similarly, we manually annotated the liver in 95 randomly chosen images from the IDEAL protocol. Next, we split the segmented IDEAL images into training ($n = 64$), validation ($n = 16$) and test ($n = 15$) sets. The overall performance of the liver segmentation is also about 94% on Dice scores. After the liver has been identified by the 2D U-net model on each slice for the two imaging protocols, we applied a 2D CNN Residual Neural Network (2D-CNN-ResNet) model using two steps on the segmented liver[87]. From the 4,616 individuals with true PDFF values, quantified by Perspectum Diagnostics from GRE imaging, we selected 4,569 individuals with a full set of ten standard liver segmentation images. We then split them into training, validation and test datasets. The 2D-CNN-ResNet model was trained and validated on 3,500 participants and tested on the remaining 1,069 participants. For the remaining 5,477 individuals from the GRE protocol, we used the CNN model developed here to predict PDFF. We then applied this 2D-CNN-ResNet model to estimate the PDFF value of participants from the IDEAL protocol. Based on these overlapping samples ($n = 1,491$) with true PDFF value derived from the first step, the 2D-CNN-ResNet model was trained ($n = 952$), validated ($n = 238$) and tested ($n = 301$). PDFF for the remaining 34,351 participants with only IDEAL imaging was then inferred using this CNN model. Inferred PDFF had a Pearson correlation coefficient of 0.976 and 0.984 in the validation and testing datasets, respectively. We also measured true PDFF values (Extended Data Fig. 8). This will be called the UKBB MRI-PDFF dataset, which after accounting for genetic missingness ($n = 1,151$) totaled $n = 43,293$ (Supplementary Table 16). A second UKBB MRI-PDFF dataset included only EUR participants and totaled $n = 41,834$ (Supplementary Table 16).

eMERGE

The eMERGE NAFLD cohort (n_{cases} = 1,106; $n_{controls}$ = 8,571) was previously described[88] and summary statistics are available at https://www.ebi.ac.uk/gwas/studies/GCST008468. EAFs were not available and were estimated using UKBB EURs.

FinnGen

We used FinnGen data freeze 4 summary statistics from https://www.finngen.fi/fi (n = 651 NAFLD cases, 176,248 controls).

MGI

MGI is a hospital-based cohort of patients seen at Michigan Medicine (Ann Arbor, MI). The MGI cohort was previously described[89]. NAFLD cases were identified by ICD-9 571.8 or ICD-10 K76.0, and HCC cases were identified by ICD-9 155.0 or ICD-10 C22.0. Cirrhosis was defined by ICD-9 571.2 or 571.5 or 571.6, or ICD-10 K70.2–4 or K74.x or K71.7 or NLP (which has been previously described)[89]. Characteristics of the included EUR ancestry participants are shown in Supplementary Table 13.

GWAS and meta-analysis

We carried out a GWAS of autosomal variants, assuming additive effects, in each of the nine GOLD cohorts separately. The analyses were corrected for age, age^2, sex, alcoholic drinks and principal components (PCs) or admixture. Sensitivity analyses by sex, study and ancestry did not show significant heterogeneity allowing us to combine the data across cohorts for all individuals with genetic data (n = 23,521). The GOLD Consortium meta-analysis was performed using a two-tailed sample size and direction of effect approach in METAL (08/28/2018 release)[90].

GWAS of autosomal variants were carried out independently in UKBB using linear mixed modeling using SAIGE (version 0.29) with binary NAFLD or inverse-normally transformed MRI-PDFF as the dependent variable using an additive genetic model[84,91]. A SNP imputation quality cutoff of >0.85 was used. The model was controlled for sex, age, age^2 and PCs 1–10.

Summary statistics from FinnGen and eMERGE studies were combined with the UKBB NAFLD, UKBB MRI-PDFF and GOLD CT steatosis analyses using sample size and direction of effect meta-analysis implemented in METAL (Extended Data Fig. 1)[90]. We call this analysis GOLDPlus. We excluded multi-allelic variants, indels, variants with minor allele frequency <0.001, variants with minor allele count <400 and variants present in less than four cohorts. We also excluded variants with $P_{het} < 0.05$ and opposing directionality across studies simultaneously. The $P < 5.00 \times 10^{-8}$ was considered genome-wide significant. Given the multi-ethnic nature of the analysis, we identified independent loci using 500-kb flanking criteria from the lowest P value associated variant. To ascertain independent signals, we also performed a direct conditional analysis for all our top hits using the UKBB multi-ethnic cohort. To perform conditional analysis, we added the genetic dosage of the loci to the other covariates (age, age^2, sex and PCs 1–10) of SAIGE step 1 and reran the GWAS. SNP-specific annotation information was obtained from ANNOVAR.

Ancestry- and sex-specific analyses in the GOLD Consortium

To assess ancestry-specific differences, we conducted a meta-analysis in the GOLD Consortium for each ancestry (EUR, AFR, HIS and CHN) separately and all ancestries together using METAL (Supplementary Table 6). Additionally, we conducted separate GWAS in men and women in the GOLD Consortium and meta-analyzed the GWAS using METAL. Sex-specific GWAS analyses were controlled for age, age^2, number of alcoholic drinks per week and PCs 1–10. Cochran's Q test was used to assess the observed heterogeneity, and the I^2 metric was used for quantification. A Cochran's Q test $P < 2.00 \times 10^{-4}$ was considered significant.

GWAS analysis stratified by alcohol use

Using the UKBB MRI-PDFF data, we performed alcohol-specific GWAS of heavy and light drinkers. We identified heavy drinkers as ≥14 drinks consumed per week for males or ≥7 drinks a week for females (n = 21,356) and light drinkers as ≤1 drinks consumed per week for males and females (n = 9,871; Supplementary Table 11). The UKBB MRI-PDFF GWAS was carried out as described above. A meta-analysis of the heavy and light drinkers was performed using METAL to assess the heterogeneity[90].

DEPICT analyses

DEPICT provides details regarding GWAS-prioritized tissues, genes and pathways across cells and tissues[23]. Enrichment was considered statistically significant at a FDR $P < 0.05$ (Extended Data Fig. 5).

Previously published NAFLD/steatosis variants

We evaluated the effects of previously reported NAFLD/steatosis variants in GOLDPlus (Supplementary Table 19). A literature search was conducted for NAFLD and steatosis GWAS in PubMed, and genome-wide significant variants were identified[8,12,16,27,92–102]. Variants that were independent of the GOLDPlus genome-wide significant variants (500 kb flanking criteria from the lowest P-value-associated variant) were assessed.

PheWAS

We used publicly available UKBB GWAS data from the Neale Lab (http://www.nealelab.is/uk-biobank/) to perform a single-variant PheWAS of the NAFLD risk-increasing alleles with related phenotypes (Fig. 2)[84,103]. Associations were considered significant with $P < 0.05$. We created PRSs from the subgroups identified in Fig. 2 using betas from GOLD as weights and carried out association analyses with cardiometabolic and GI traits in UKBB. Associations were adjusted for age, age^2, sex and PCs 1–10. Outcomes are reported in s.d. for continuous traits and log OR for disease outcomes of the top tertile or quartile (for low output only due to the distribution of scores) versus the lowest tertile or value of zero of PRS, respectively.

PheWAS Manhattan plot of NAFLD PRS

For each ICD code in MGI, we fit a logistic regression model using Firth's logistic regression. Every model included the ICD code as binary outcome variable and PRS built

on the GOLD-weighted score of the 17 genome-wide SNPs, age, sex and the first ten PCs as predictors. Associations were considered suggestive at $\alpha = 0.05$ and significant at Bonferroni-adjusted $\alpha = 3.02 \times 10^{-5}$ P values (Fig. 5). The associations were plotted using R version 4.0.2.

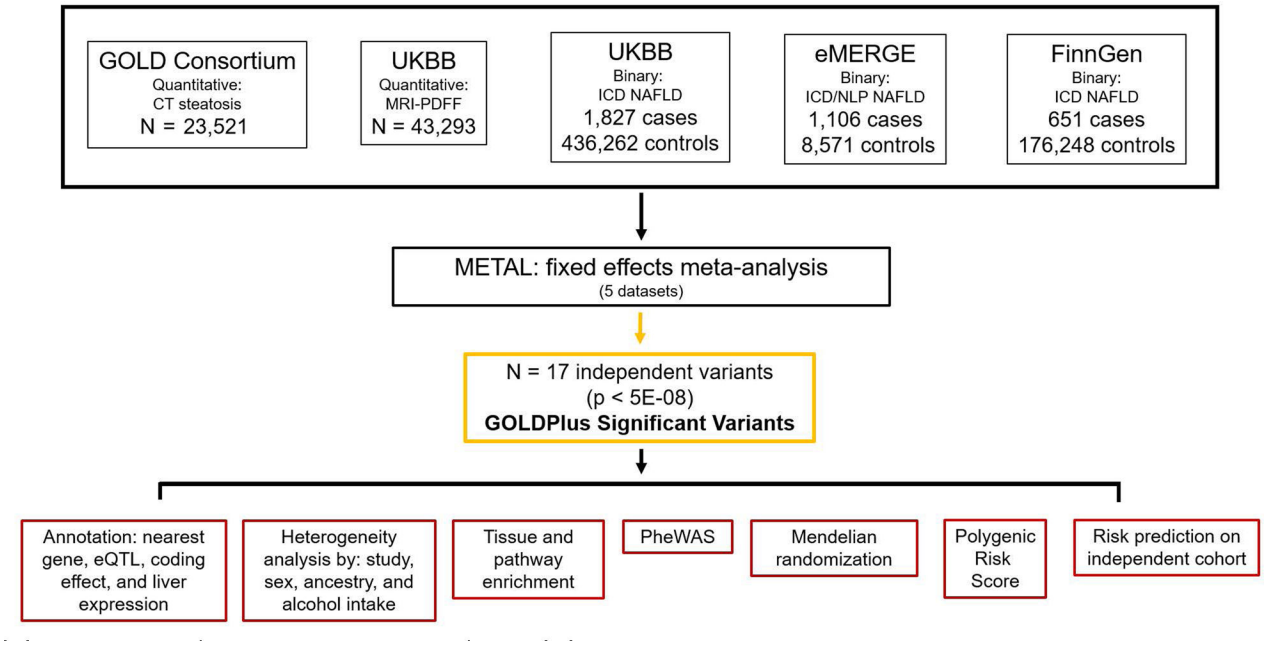
PRS and NAFLD risk factors

We created a PRS using the liver fat-increasing variants ($n = 17$) from the GOLDPlus meta-analysis. The PRS was based on a weighted sum of dosage of the NAFLD-associated single variants. The β value of each allele (from the GOLD Consortium) was used to weigh the PRS. The predictive power of the PRS was assessed on NAFLD, cirrhosis and HCC cohorts in MGI EUR ancestry samples (Fig. 6 and Supplementary Table 9). PRSs were defined as inverse-normally transformed rank units or as percentiles. Analyses were adjusted for age, age^2, sex and PCs 1–10.

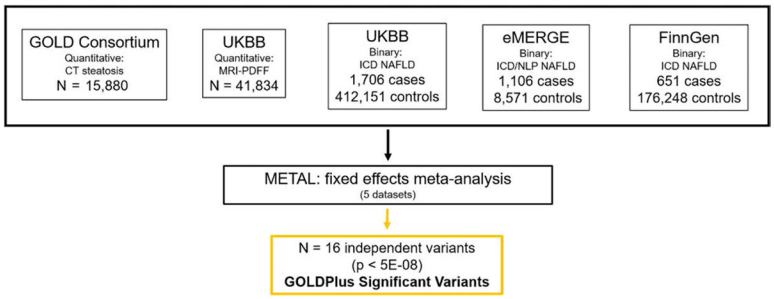
MR

We performed a two-sample MR, implemented in R version 3.5.1 using 'TwoSampleMR' (version 0.5.5)[104]. For the analysis, we used the variant-NAFLD effect estimates from the GOLD Consortium (betas are required for MR and the GOLD Consortium data had the highest quality measures of hepatic steatosis in the population-based cohorts). We calculated F-statistics[105] for each variant, and only those having an F-statistic >10 were included in the MR analysis (Supplementary Table 14)[26]. MR was performed using the resulting variants as the exposure and related publicly available (Supplementary Table 15) and UKBB GWAS (K74 fibrosis and cirrhosis of liver and I85 esophageal varices, a complication of cirrhosis) as outcomes. We also performed the reverse analysis where independent genome-wide significant ($P < 5.00 \times 10^{-8}$) variants from the aforementioned GWAS were used as exposure and the GOLD Consortium phenotype as the outcome. We applied inverse-variance weighted, penalized weighted median, weighted median, weighted mode and MR-Egger methods. Tests for heterogeneity and horizontal pleiotropy were also performed (Extended Data Figs. 6 and 7).

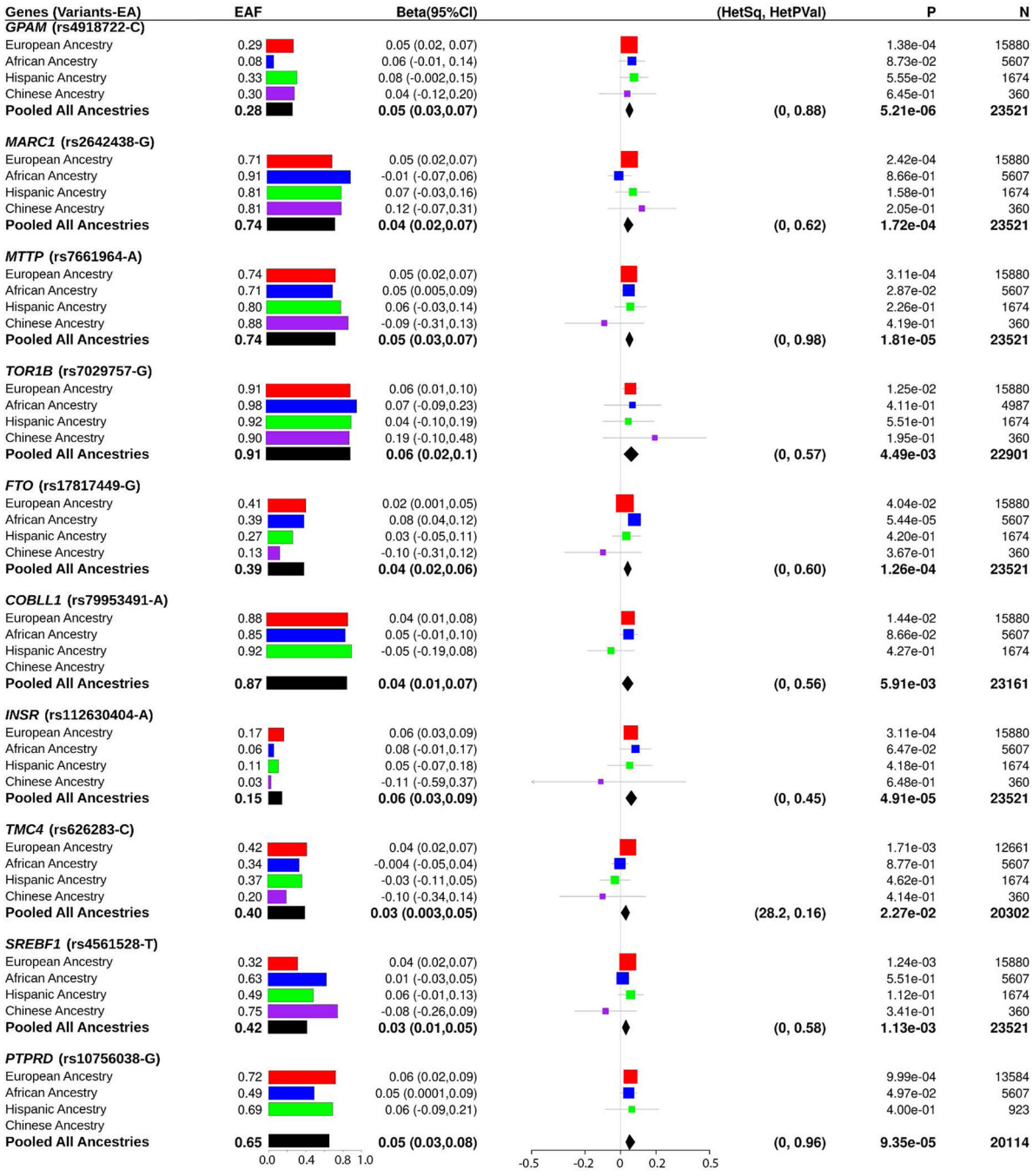
Extended Data



Extended Data Fig. 1 |. GOLDPlus NAFLD measures meta-analysis study design.



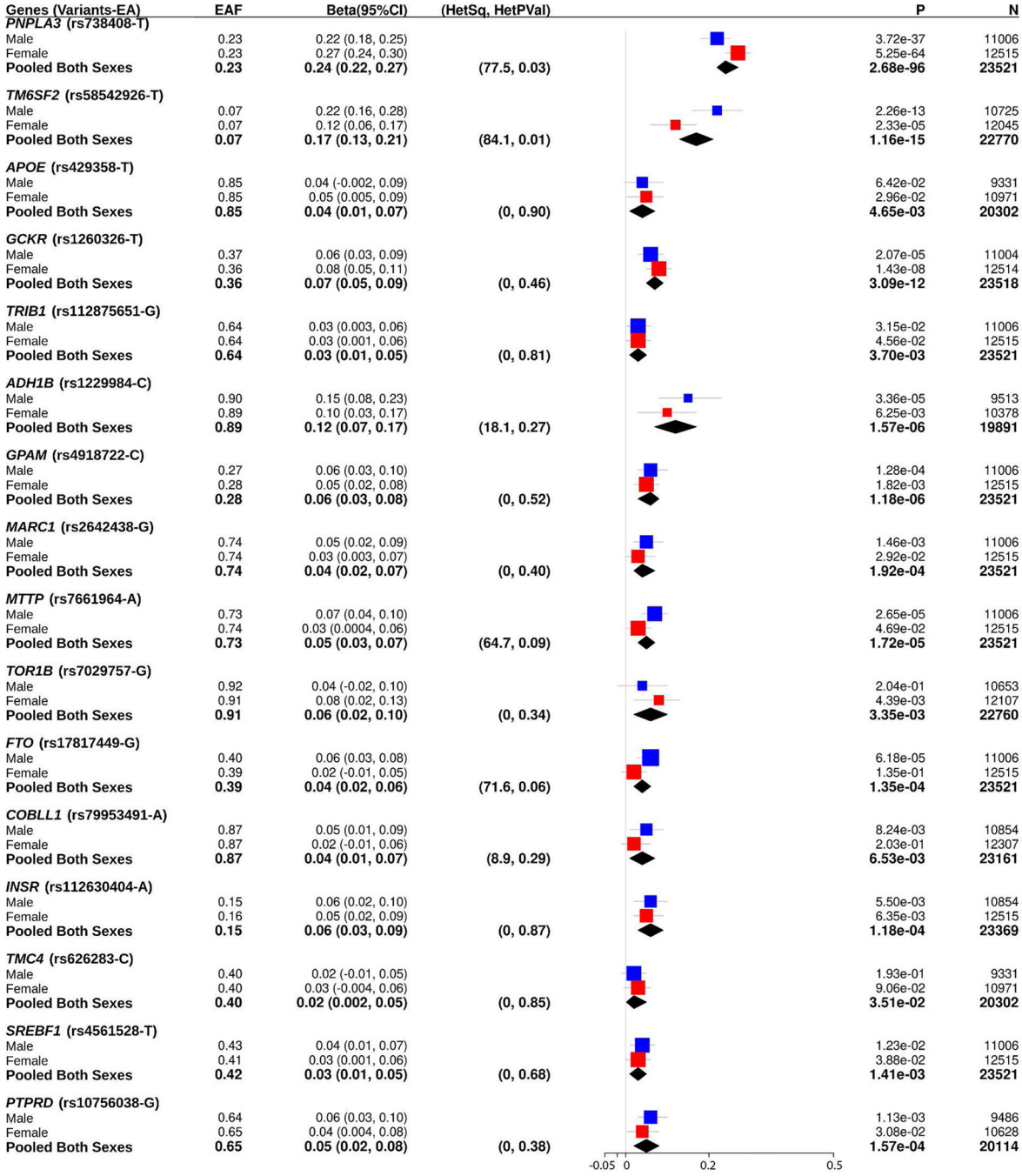
Extended Data Fig. 2 |. European GOLDPlus NAFLD measures meta-analysis schematic.



Extended Data Fig. 3 |. Characteristics of GOLDPlus genome-wide significant variants in GOLD ancestry-based cohorts.

For each variant, the characteristics are shown for the GOLD ancestry-based analysis including: associated gene, NAFLD increasing effect allele (EA), effect allele frequency (EAF), effect/beta and 95% confidence interval (CI), Cochran's Q heterogeneity I^2 metric (HetSq) and heterogeneity *P*-value (HetPVal), EA *P*-value (*P*), and sample size (*N*). Results are for meta-analysis of GOLD European ancestry (red), African ancestry (blue), Hispanic ancestry (green), Chinese ancestry (purple), and all ancestries pooled (black). The estimates of the effect sizes (Beta) and 95% confidence interval in bidirectional testing within each

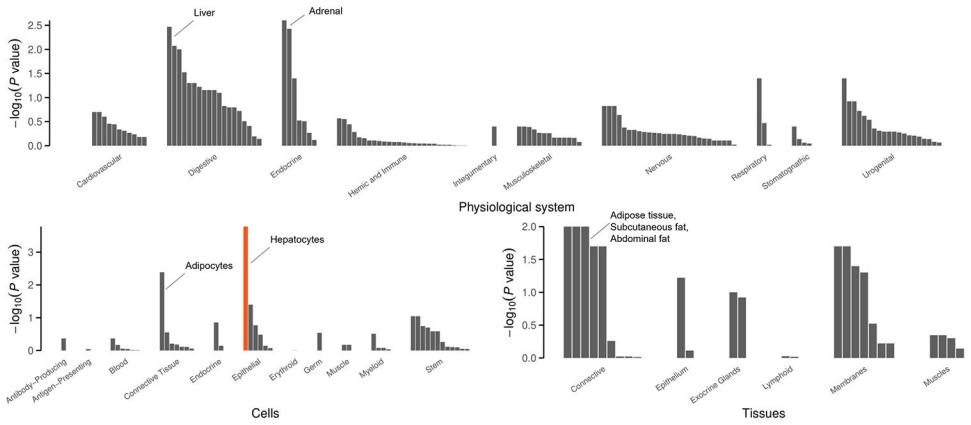
ancestry and for all the ancestries combined were shown in the forest plots. The data underlying these plots are provided as Source Data.



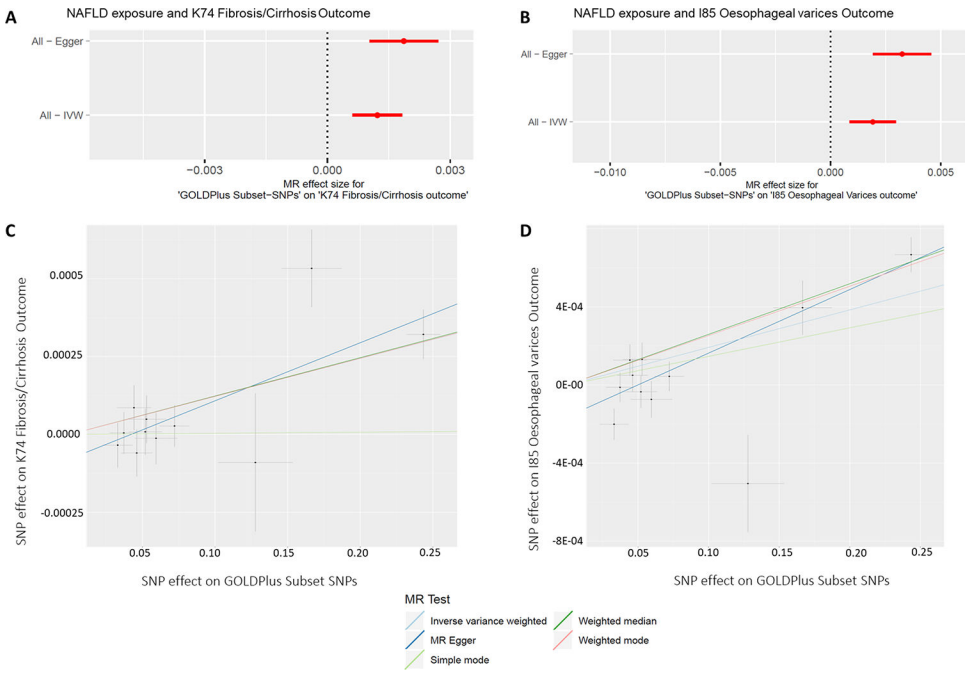
Extended Data Fig. 4 |. Characteristics of GOLDPlus genome-wide significant variants in GOLD sex-specific cohorts.

For each variant, the characteristics are shown for the GOLD sex-specific analysis including: associated gene, NAFLD increasing effect allele (EA), effect allele frequency (EAF), effect/beta and 95% confidence interval (CI), Cochran's Q heterogeneity I^2 metric (HetSq) and heterogeneity *P*-value (HetPVal), EA *P*-value (*P*), and sample size (*N*). Results are for

meta-analysis of GOLD cohort males (blue), females (red), and pooled sexes (black). The estimates of the effect sizes (Beta) and 95% confidence interval in bidirectional testing within each ancestry and for all the ancestries combined were shown in the forest plots. The data underlying these plots are provided as Source Data.

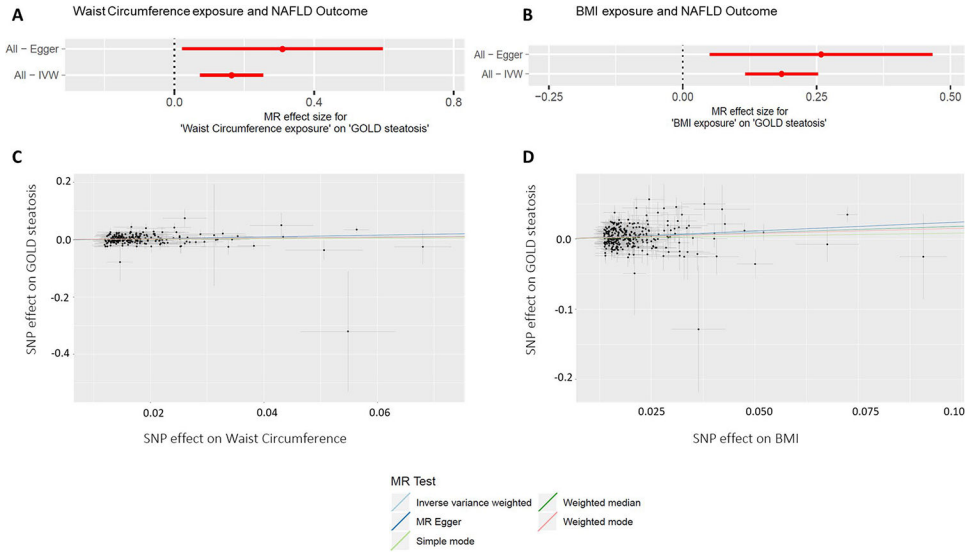


Extended Data Fig. 5 |. DEPICT analysis of biological enrichment of NAFLD associated variants. Height of the bar represents the nominal $-\log_{10}P$-value of enrichment of GWAS associated genes with physiological systems, cells, and tissues. Dark orange shading represents statistical significance at false discovery rate (FDR) < 0.05. The data underlying these plots are provided as Source Data.



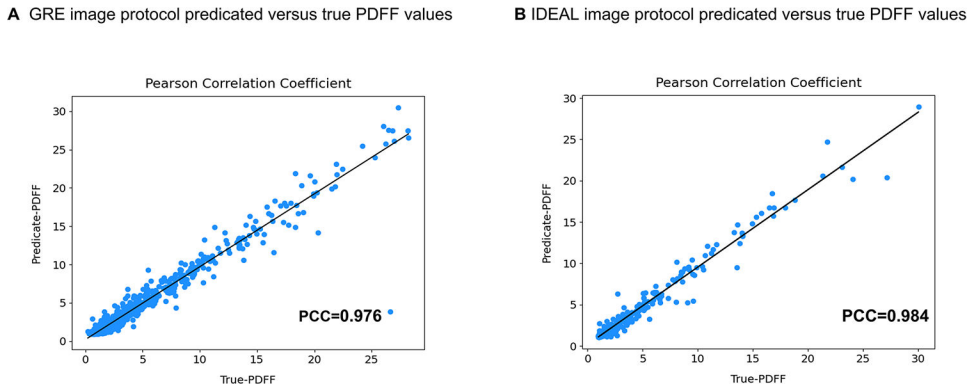
Extended Data Fig. 6 |. Two-sample Mendelian randomization analysis for casual associations between NAFLD associated variants and fibrosis/cirrhosis and esophageal varices. **a**,**b**, Data represent the effect/beta and 95% confidence intervals for the inverse variance weighted (IVW) and MR-Egger analyses for (**a**) NAFLD exposure (GOLD cohort, n =

11 instruments) and K74:fibrosis/cirrhosis outcome (UKBB) (MR-Egger *P*-value = 1.88 × 10^{-3}, IVW p-value = 8.65 × 10^{-5}) and (**b**) NAFLD exposure (GOLD cohort, *n* = 11 instruments) and I85:esophageal varices outcome (UKBB) (MR-Egger *P*-value = 9.36 × 10^{-4}, IVW *P*-value = 3.51 × 10^{-4}). **c**,**d**, The crosshairs on the plots represent the effect and 95% confidence intervals for each SNP-NAFLD or SNP-outcome association for (**c**) NAFLD exposure (GOLD cohort, *n* = 10 instruments) and K74:fibrosis/cirrhosis outcome (UKBB) and (**d**) NAFLD exposure (GOLD cohort, *n* = 10 instruments) and I85:esophageal varices outcome (UKBB). The data underlying these plots are provided as Source Data.



Extended Data Fig. 7 |. Two-sample Mendelian randomization analysis for casual associations between BMI, waist circumference associated variants and NAFLD.

a,**b**, Data are presented as effect/beta and 95% confidence intervals for MR-Egger and inverse variance weighted (IVW) methods for (**a**) waist circumference GWAS (UKBB, *n* = 217 instruments) and GOLD cohort outcome (MR-Egger *P*-value = 3.6 × 10^{-2}, IVW *P*-value = 3.71 × 10^{-4}) and (**b**) BMI GWAS (UKBB, *n* = 293 instruments) and GOLD cohort outcome (MR-Egger *P*-value = 0.02, IVW *P*-value = 1.02 × 10^{-7}). **c**,**d**, The crosshairs on the plots represent the effect/beta and 95% confidence intervals for each SNP-NAFLD or SNP-outcome association for (**c**) waist circumference GWAS (UKBB, *n* = 211 instruments) and GOLD cohort outcome and (**d**) BMI GWAS (UKBB, *n* = 283 instruments) and GOLD cohort outcome. The data underlying these plots are provided as Source Data.



Extended Data Fig. 8 |. Convolutional neural network schematic for UKBB MRI liver imaging (PCC values).

Scatter plot of predicted UKBB MRI-PDFF values versus 'true' UKBB MRI-PDFF values (as determined by Perspectum Diagnostics). **a**,**b**, Pearson correlation coefficients (PCC) are shown for (**a**) gradient echo image protocol and (**b**) IDEAL image protocol. The data underlying these plots are provided as Source Data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

AGES was funded by the National Institutes of Health (NIH; contracts N01-AG-1-2100 and HHSN271201200022C), the NIA Intramural Research Program, Hjartavernd (the Icelandic Heart Association) and the Althingi (the Icelandic Parliament). Support for FamHS was provided by the National Heart, Lung and Blood Institute (NHLBI; grants R01 HL087700 and R01 HL117078) and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK; grant R01 DK089256 to M.A.P.). FHS is conducted and supported by the NHLBI in collaboration with Boston University (contracts N01-HC-25195, HHSN2682015000001I and 75N92019D00031). Funding for SHARe Affymetrix genotyping was provided by NHLBI (contract N02-HL64278). SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University. The Old Order Amish liver phenotyping is supported by NIH grants and contracts (U01 HL072515 and P30 DK72488) and analysis methods by U01 HL137181 (to J.R.O.). Support for the GENOA study was provided by the NIH (grants HL085571 to P.A.P. and HL087660 to S.L.R.K.) and NHLBI (HL100245). Support for the IRASFS was provided by the NHLBI (grants R01 HL060944, R01 HL061019, R01 HL060919, R01 HL060894 and R01 HL061210 to X.G., D.W.B., J.M.N., J.I.R., L.E.W. and N.D.P.). Genotyping and analysis were supported by NIDDK (grants DK085175 and R01 DK118062). JHS is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the NHLBI and the National Institute on Minority Health and Health Disparities (NIMHD). We also thank the staff and participants of the JHS. MESA and the MESA SHARe projects are conducted and supported by the NHLBI in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079 and UL1-TR-001420, UL1TR001881, DK063491 and R01HL105756. Funding for SHARe genotyping was provided by NHLBI (contract N02-HL-64278). Genotyping was performed at Affymetrix and the Broad Institute of Harvard and MIT (Boston, MA) using the Affymetrix Genome-Wide Human SNP Array 6.0. We thank the other investigators, the staff and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutes can be found at http://www.mesa-nhlbi.org. L.F.B. was supported by R01 HL071739 for all measures of NAFLD in MESA. OOA studies are supported by grants and contracts from NIH, including U01 HL072515, U01 HL84756, U01 HL137181 and P30 DK72488. We acknowledge the MGI participants, Precision Health at the University of Michigan, the University of Michigan Medical School Central Biorepository and the University of Michigan Advanced Genomics Core for providing data and specimen storage, management, processing and distribution services. We also acknowledge the Center for Statistical Genetics

in the Department of Biostatistics at the School of Public Health for genotype data curation, imputation and management in support of the research reported in this publication. COPDGene is supported by NHLBI (U01 HL089897 and U01 HL089856) as well as through contributions made to an industry advisory board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion. Liver fat measures in COPDGene were gathered under HL122464. Analyses in the UKBB were done under approved project 18120 (to E.K.S.). E.K.S., Y.C., A.K., X.D., A.O. and B.D.H. are supported by NIH (grants R01 DK106621 and R01 DK107904 to E.K.S.) and The University of Michigan Department of Internal Medicine. N.D.P. and E.K.S. are supported by NIH (grants R01 DK128871 to N.D.P. and E.K.S.; R01DK131787 to E.K.S.). V.L.C. was supported in part by an American Association for the Study of Liver Disease Clinical, Translational and Outcomes Research Award. We acknowledge the participants and investigators of the FinnGen study. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the NHLBI; the National Institutes of Health; the US Department of Health and Human Services; Framingham Heart Study or Boston University.

Data availability

Meta-analysis results from this study are available at http://www.med.umich.edu/spelioteslab/ and at GWAS Catalog (GCP ID: GCP000662). GOLD Consortium and MGI individual-level data are governed by patient privacy requirements and available to those having the mandatory IRB approvals. The eMERGE NAFLD cohort was previously described, and summary statistics are publicly available (https://www.ebi.ac.uk/gwas/studies/GCST008468). FinnGen data freeze 4 summary statistics are publicly available (https://www.finngen.fi/fi). UKBB genomic and phenotypic data supporting this publication are available upon application (https://ukbiobank.ac.uk). Otherwise, all data used to generate figures can be found in supplementary tables or in the above publicly available datasets. Source data are provided with this paper.

References

1. Lazo M et al. Prevalence of nonalcoholic fatty liver disease in the United States: the third National Health and Nutrition Examination Survey, 1988–1994. Am. J. Epidemiol. 178, 38–45 (2013). [PubMed: 23703888]
2. Portillo Sanchez P et al. High prevalence of nonalcoholic fatty liver disease in patients with type 2 diabetes mellitus and normal plasma aminotransferase levels. J. Clin. Endocrinol. Metab. 100, 2231–2238 (2015). [PubMed: 25885947]
3. Dongiovanni P et al. Causal relationship of hepatic fat with liver damage and insulin resistance in nonalcoholic fatty liver. J. Intern. Med. 283, 356–370 (2018). [PubMed: 29280273]
4. Lauridsen BK et al. Liver fat content, non-alcoholic fatty liver disease, and ischaemic heart disease: Mendelian randomization and meta-analysis of 279 013 individuals. Eur. Heart J. 39, 385–393 (2018). [PubMed: 29228164]
5. Stender S et al. Adiposity amplifies the genetic risk of fatty liver disease conferred by multiple loci. Nat. Genet. 49, 842–847 (2017). [PubMed: 28436986]
6. Liu Z et al. Causal relationships between NAFLD, T2D and obesity have implications for disease subphenotyping. J. Hepatol. 73, 263–276 (2020). [PubMed: 32165250]
7. Bianco C et al. Non-invasive stratification of hepatocellular carcinoma risk in non-alcoholic fatty liver using polygenic risk scores. J. Hepatol. 74, 775–782 (2021). [PubMed: 33248170]
8. Parisinos CA et al. Genome-wide and Mendelian randomisation studies of liver MRI yield insights into the pathogenesis of steatohepatitis. J. Hepatol. 73, 241–251 (2020). [PubMed: 32247823]
9. Crespo J et al. Are there predictive factors of severe liver fibrosis in morbidly obese patients with non-alcoholic steatohepatitis? Obes. Surg. 11, 254–257 (2001). [PubMed: 11433896]
10. Younossi ZM et al. The economic and clinical burden of nonalcoholic fatty liver disease in the United States and Europe. Hepatology 64, 1577–1586 (2016). [PubMed: 27543837]
11. Romeo S et al. Genetic variation in *PNPLA3* confers susceptibility to nonalcoholic fatty liver disease. Nat. Genet. 40, 1461–1465 (2008). [PubMed: 18820647]

12. Speliotes EK et al. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. PLoS Genet. 7, e1001324 (2011). [PubMed: 21423719]
13. Luukkonen PK et al. *MARC1* variant rs2642438 increases hepatic phosphatidylcholines and decreases severity of non-alcoholic fatty liver disease in humans. J. Hepatol. 73, 725–726 (2020). [PubMed: 32471727]
14. Jamialahmadi O et al. Exome-wide association study on alanine aminotransferase identifies sequence variants in the *GPAM* and *APOE* associated with fatty liver disease. Gastroenterology 160, 1634–1646 (2021). [PubMed: 33347879]
15. Kitamoto A et al. Association of polymorphisms in *GCKR* and *TRIB1* with nonalcoholic fatty liver disease and metabolic syndrome traits. Endocr. J. 61, 683–689 (2014). [PubMed: 24785259]
16. Mancina RM et al. The *MBOAT7-TMC4* variant rs641738 increases risk of nonalcoholic fatty liver disease in individuals of European descent. Gastroenterology 150, 1219–1230 (2016). [PubMed: 26850495]
17. Nakajima S et al. Polymorphism of receptor-type tyrosine-protein phosphatase δ gene in the development of non-alcoholic fatty liver disease. J. Gastroenterol. Hepatol. 33, 283–290 (2018). [PubMed: 28497593]
18. Palmer ND et al. Allele-specific variation at *APOE* increases nonalcoholic fatty liver disease and obesity but decreases risk of Alzheimer's disease and myocardial infarction. Hum. Mol. Genet. 30, 1443–1456 (2021). [PubMed: 33856023]
19. Vilar-Gomez E et al. *ADH1B*2* is associated with reduced severity of nonalcoholic fatty liver disease in adults, independent of alcohol consumption. Gastroenterology 159, 929–943 (2020). [PubMed: 32454036]
20. Zheng W et al. *MTP* −493G>T polymorphism and susceptibility to nonalcoholic fatty liver disease: a meta-analysis. DNA Cell Biol. 33, 361–369 (2014). [PubMed: 24588800]
21. Middleton MS et al. Agreement between magnetic resonance imaging proton density fat fraction measurements and pathologist-assigned steatosis grades of liver biopsies from adults with nonalcoholic steatohepatitis. Gastroenterology 153, 753–761 (2017). [PubMed: 28624576]
22. Saadeh S et al. The utility of radiological imaging in nonalcoholic fatty liver disease. Gastroenterology 123, 745–750 (2002). [PubMed: 12198701]
23. Pers TH et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nat. Commun. 6, 5890 (2015). [PubMed: 25597830]
24. Kahali B et al. A noncoding variant near *PPP1R3B* promotes liver glycogen storage and MetS, but protects against myocardial infarction. J. Clin. Endocrinol. Metab. 106, 372–387 (2021). [PubMed: 33231259]
25. Liberzon A et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740 (2011). [PubMed: 21546393]
26. Lawlor DA et al. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat. Med. 27, 1133–1163 (2008). [PubMed: 17886233]
27. Chen VL et al. Genome-wide association study of serum liver enzymes implicates diverse metabolic and liver pathology. Nat. Commun. 12, 816 (2021). [PubMed: 33547301]
28. Emdin CA et al. Association of genetic variation with cirrhosis: a multi-trait genome-wide association and gene-environment interaction study. Gastroenterology 160, 1620–1633 (2021). [PubMed: 33310085]
29. Sveinbjornsson G et al. Multiomics study of nonalcoholic fatty liver disease. Nat. Genet. 54, 1652–1663 (2022). [PubMed: 36280732]
30. Vujkovic M et al. A multiancestry genome-wide association study of unexplained chronic ALT elevation as a proxy for nonalcoholic fatty liver disease with histological and radiological validation. Nat. Genet. 54, 761–771 (2022). [PubMed: 35654975]
31. Chambers JC et al. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. Nat. Genet. 43, 1131–1138 (2011). [PubMed: 22001757]
32. Kathiresan S et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat. Genet. 40, 189–197 (2008). [PubMed: 18193044]

33. Beer NL et al. The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. Hum. Mol. Genet. 18, 4081–4088 (2009). [PubMed: 19643913]
34. Ishizuka Y et al. TRIB1 downregulates hepatic lipogenesis and glycogenesis via multiple molecular interactions. J. Mol. Endocrinol. 52, 145–158 (2014). [PubMed: 24389359]
35. Bauer RC et al. Tribbles-1 regulates hepatic lipogenesis through posttranscriptional regulation of C/EBPalpha. J. Clin. Invest. 125, 3809–3818 (2015). [PubMed: 26348894]
36. Agius L Hormonal and metabolite regulation of hepatic glucokinase. Annu. Rev. Nutr. 36, 389–415 (2016). [PubMed: 27146014]
37. Janssen MC et al. Symptomatic lipid storage in carriers for the *PNPLA2* gene. Eur. J. Hum. Genet. 21, 807–815 (2013). [PubMed: 23232698]
38. Steneberg P et al. Hyperinsulinemia enhances hepatic expression of the fatty acid transporter Cd36 and provokes hepatosteatosis and hepatic insulin resistance. J. Biol. Chem. 290, 19034–19043 (2015). [PubMed: 26085100]
39. Ipsen DH, Lykkesfeldt J & Tveden-Nyborg P Molecular mechanisms of hepatic lipid accumulation in non-alcoholic fatty liver disease. Cell. Mol. Life Sci. 75, 3313–3327 (2018). [PubMed: 29936596]
40. Popineau L et al. Novel Grb14-mediated cross talk between insulin and p62/Nrf2 pathways regulates liver lipogenesis and selective insulin resistance. Mol. Cell. Biol. 36, 2168–2181 (2016). [PubMed: 27215388]
41. Cooney GJ et al. Improved glucose homeostasis and enhanced insulin signalling in Grb14-deficient mice. EMBO J. 23, 582–593 (2004). [PubMed: 14749734]
42. Michael MD et al. Loss of insulin signaling in hepatocytes leads to severe insulin resistance and progressive hepatic dysfunction. Mol. Cell 6, 87–97 (2000). [PubMed: 10949030]
43. Sirwi A & Hussain MM Lipid transfer proteins in the assembly of apoB-containing lipoproteins. J. Lipid Res. 59, 1094–1102 (2018). [PubMed: 29650752]
44. Adam MP et al. (eds.). GeneReviews((R)) (University of Washington, 1993).
45. Consortium GTEx. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330 (2020). [PubMed: 32913098]
46. Polimanti R & Gelernter J *ADH1B*: from alcoholism, natural selection, and cancer to the human phenome. Am. J. Med. Genet. B Neuropsychiatr. Genet. 177, 113–125 (2018). [PubMed: 28349588]
47. Gu S et al. Recent selection on a class I ADH locus distinguishes Southwest Asian populations including Ashkenazi Jews. Genes (Basel) 9, 452 (2018). [PubMed: 30205534]
48. Macgregor S et al. Associations of *ADH* and *ALDH2* gene variation with self report alcohol reactions, consumption and dependence: an integrated analysis. Hum. Mol. Genet. 18, 580–593 (2009). [PubMed: 18996923]
49. Muenter MD, Perry HO & Ludwig J Chronic vitamin A intoxication in adults. Hepatic, neurologic and dermatologic complications. Am. J. Med. 50, 129–136 (1971). [PubMed: 4099655]
50. Shin JY et al. Nuclear envelope-localized torsinA-LAP1 complex regulates hepatic VLDL secretion and steatosis. J. Clin. Invest. 129, 4885–4900 (2019). [PubMed: 31408437]
51. Innes H et al. Genome-wide association study for alcohol-related cirrhosis identifies risk loci in *MARC1* and *HNRNPUL1*. Gastroenterology 159, 1276–1289.e7 (2020). [PubMed: 32561361]
52. Xia M et al. Hepatic deletion of *Mboat7 (Lpiat1)* causes activation of SREBP-1c and fatty liver. J. Lipid Res. 62, 100031 (2021). [PubMed: 32859645]
53. Landgraf K et al. *FTO* obesity risk variants are linked to adipocyte *IRX3* expression and BMI of children—relevance of *FTO* variants to defend body weight in lean children? PLoS ONE 11, e0161739 (2016). [PubMed: 27560134]
54. Kozlitina J et al. Exome-wide association study identifies a *TM6SF2* variant that confers susceptibility to nonalcoholic fatty liver disease. Nat. Genet. 46, 352–356 (2014). [PubMed: 24531328]
55. Wang Y et al. PNPLA3, CGI-58, and inhibition of hepatic triglyceride hydrolysis in mice. Hepatology 69, 2427–2441 (2019). [PubMed: 30802989]

56. Morton AM et al. Apolipoproteins E and CIII interact to regulate HDL metabolism and coronary heart disease risk. JCI Insight 3, e98045 (2018). [PubMed: 29467335]
57. Abul-Husn NS et al. A protein-truncating *HSD17B13* variant and protection from chronic liver disease. N. Engl. J. Med. 378, 1096–1106 (2018). [PubMed: 29562163]
58. Fox CS et al. Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. PLoS Genet. 8, e1002695 (2012). [PubMed: 22589738]
59. Hatters DM, Peters-Libeu CA & Weisgraber KH Apolipoprotein E structure: insights into function. Trends Biochem. Sci. 31, 445–454 (2006). [PubMed: 16820298]
60. Mahdessian H et al. TM6SF2 is a regulator of liver fat metabolism influencing triglyceride secretion and hepatic lipid droplet content. Proc. Natl Acad. Sci. USA 111, 8913–8918 (2014). [PubMed: 24927523]
61. BasuRay S et al. Accumulation of PNPLA3 on lipid droplets is the basis of associated hepatic steatosis. Proc. Natl Acad. Sci. USA 116, 9521–9526 (2019). [PubMed: 31019090]
62. Rondinone CM et al. Protein tyrosine phosphatase 1B reduction regulates adiposity and expression of genes involved in lipogenesis. Diabetes 51, 2405–2411 (2002). [PubMed: 12145151]
63. Zou Y et al. IRX3 promotes the browning of white adipocytes and its rare variants are associated with human obesity risk. EBioMedicine 24, 64–75 (2017). [PubMed: 28988979]
64. Zeng H et al. CD36 promotes de novo lipogenesis in hepatocytes through INSIG2-dependent SREBP1 processing. Mol. Metab. 57, 101428 (2022). [PubMed: 34974159]
65. Cignarelli A et al. Insulin and insulin receptors in adipose tissue development. Int. J. Mol. Sci. 20, 759 (2019). [PubMed: 30754657]
66. Ong KT et al. Adipose triglyceride lipase is a major hepatic lipase that regulates triacylglycerol turnover and fatty acid signaling and partitioning. Hepatology 53, 116–126 (2011). [PubMed: 20967758]
67. Morales LD et al. Further evidence supporting a potential role for ADH1B in obesity. Sci. Rep. 11, 1932 (2021). [PubMed: 33479282]
68. Tanaka Y et al. LPIAT1/MBOAT7 depletion increases triglyceride synthesis fueled by high phosphatidylinositol turnover. Gut 70, 180–193 (2021). [PubMed: 32253259]
69. Neschen S et al. Prevention of hepatic steatosis and hepatic insulin resistance in mitochondrial acyl-CoA:glycerol-sn-3-phosphate acyltransferase 1 knockout mice. Cell Metab. 2, 55–65 (2005). [PubMed: 16054099]
70. Linden D et al. Liver-directed overexpression of mitochondrial glycerol-3-phosphate acyltransferase results in hepatic steatosis, increased triacylglycerol secretion and reduced fatty acid oxidation. FASEB J. 20, 434–443 (2006). [PubMed: 16507761]
71. Klein JM et al. The mitochondrial amidoxime-reducing component (mARC1) is a novel signal-anchored protein of the outer mitochondrial membrane. J. Biol. Chem. 287, 42795–42803 (2012). [PubMed: 23086957]
72. Hussain MM et al. Multiple functions of microsomal triglyceride transfer protein. Nutr. Metab. (Lond.) 9, 14 (2012). [PubMed: 22353470]
73. Fernandes Silva L et al. An intronic variant in the *GCKR* gene is associated with multiple lipids. Sci. Rep. 9, 10240 (2019). [PubMed: 31308433]
74. Douvris A et al. Functional analysis of the *TRIB1* associated locus linked to plasma triglycerides and coronary artery disease. J. Am. Heart Assoc. 3, e000884 (2014). [PubMed: 24895164]
75. Harris TB et al. Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. Am. J. Epidemiol. 165, 1076–1087 (2007). [PubMed: 17351290]
76. Regan EA et al. Genetic epidemiology of COPD (COPDGene) study design. COPD 7, 32–43 (2010). [PubMed: 20214461]
77. Carr JJ et al. Calcified coronary artery plaque measurement with cardiac CT in population-based studies: standardized protocol of Multi-Ethnic Study of Atherosclerosis (MESA) and Coronary Artery Risk Development in Young Adults (CARDIA) study. Radiology 234, 35–43 (2005). [PubMed: 15618373]
78. Speliotes EK et al. Liver fat is reproducibly measured using computed tomography in the Framingham Heart Study. J. Gastroenterol. Hepatol. 23, 894–899 (2008). [PubMed: 18565021]

79. Daniels PR et al. Familial aggregation of hypertension treatment and control in the Genetic Epidemiology Network of Arteriopathy (GENOA) study. Am. J. Med. 116, 676–681 (2004). [PubMed: 15121494]
80. Palmer ND et al. Genetic variants associated with quantitative glucose homeostasis traits translate to type 2 diabetes in Mexican Americans: the GUARDIAN (Genetics Underlying Diabetes in Hispanics) Consortium. Diabetes 64, 1853–1866 (2015). [PubMed: 25524916]
81. Liu J et al. Fatty liver, abdominal adipose tissue and atherosclerotic calcification in African Americans: the Jackson Heart Study. Atherosclerosis 224, 521–525 (2012). [PubMed: 22902209]
82. Kramer H et al. Racial/ethnic differences in hypertension and hypertension treatment and control in the multi-ethnic study of atherosclerosis (MESA). Am. J. Hypertens. 17, 963–970 (2004). [PubMed: 15485761]
83. Rampersaud E et al. The association of coronary artery calcification and carotid artery intima-media thickness with distinct, traditional coronary artery disease risk factors in asymptomatic adults. Am. J. Epidemiol. 168, 1016–1023 (2008). [PubMed: 18805900]
84. Canela-Xandri O, Rawlik K & Tenesa A An atlas of genetic associations in UK Biobank. Nat. Genet. 50, 1593–1599 (2018). [PubMed: 30349118]
85. Ronneberger O, Fischer P & Brox T U-Net: Convolutional Networks for Biomedical Image Segmentation (Springer International Publishing, 2015).
86. Yushkevich PA et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31, 1116–1128 (2006). [PubMed: 16545965]
87. He K, Zhang X, Ren S & Sun J Deep residual learning for image recognition. In Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 770–778 (IEEE, 2016).
88. Namjou B et al. GWAS and enrichment analyses of non-alcoholic fatty liver disease identify new trait-associated genes and pathways across eMERGE Network. BMC Med. 17, 135 (2019). [PubMed: 31311600]
89. Chen VL et al. Genetic variants that associate with cirrhosis have pleiotropic effects on human traits. Liver Int. 40, 405–415 (2020). [PubMed: 31815349]
90. Willer CJ, Li Y & Abecasis GR METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191 (2010). [PubMed: 20616382]
91. Zhou W et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. 50, 1335–1341 (2018). [PubMed: 30104761]
92. Dongiovanni P et al. Genetic variants regulating insulin receptor signalling are associated with the severity of liver damage in patients with non-alcoholic fatty liver disease. Gut 59, 267–273 (2010). [PubMed: 20176643]
93. Feitosa MF et al. The *ERLIN1-CHUK-CWF19L1* gene cluster influences liver fat deposition and hepatic inflammation in the NHLBI Family Heart Study. Atherosclerosis 228, 175–180 (2013). [PubMed: 23477746]
94. Chalasani N et al. Genome-wide association study identifies variants associated with histologic features of nonalcoholic fatty liver disease. Gastroenterology 139, 1567–1576 (2010). [PubMed: 20708005]
95. Eslam M et al. Interferon-λ rs12979860 genotype and liver fibrosis in viral and non-viral chronic liver disease. Nat. Commun. 6, 6422 (2015). [PubMed: 25740255]
96. Wiedmann S et al. Genetic variants within the *LPIN1* gene, encoding lipin, are influencing phenotypes of the metabolic syndrome in humans. Diabetes 57, 209–217 (2008). [PubMed: 17940119]
97. Shang XR et al. GWAS-identified common variants with nonalcoholic fatty liver disease in Chinese children. J. Pediatr. Gastroenterol. Nutr. 60, 669–674 (2015). [PubMed: 25522307]
98. Petta S et al. IL28B and PNPLA3 polymorphisms affect histological liver damage in patients with non-alcoholic fatty liver disease. J. Hepatol. 56, 1356–1362 (2012). [PubMed: 22314430]
99. Kitamoto T et al. Genome-wide scan revealed that polymorphisms in the *PNPLA3*, *SAMM50*, and *PARVB* genes are associated with development and progression of nonalcoholic fatty liver disease in Japan. Hum. Genet. 132, 783–792 (2013). [PubMed: 23535911]

100. Anstee QM et al. Genome-wide association study of non-alcoholic fatty liver and steatohepatitis in a histologically characterised cohort. J. Hepatol. 73, 505–515 (2020). [PubMed: 32298765]

101. Ma Y et al. 17-β hydroxysteroid dehydrogenase 13 is a hepatic retinol dehydrogenase associated with histological features of nonalcoholic fatty liver disease. Hepatology 69, 1504–1519 (2019). [PubMed: 30415504]

102. Park SL et al. Genome-wide association study of liver fat: the Multiethnic Cohort Adiposity Phenotype Study. Hepatol. Commun. 4, 1112–1123 (2020). [PubMed: 32766472]

103. Martin AR et al. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. 51, 584–591 (2019). [PubMed: 30926966]

104. Hemani G et al. The MR-base platform supports systematic causal inference across the human phenome. eLife 7, e34408 (2018). [PubMed: 29846171]

105. Bowden J et al. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I^2 statistic. Int. J. Epidemiol. 45, 1961–1974 (2016). [PubMed: 27616674]

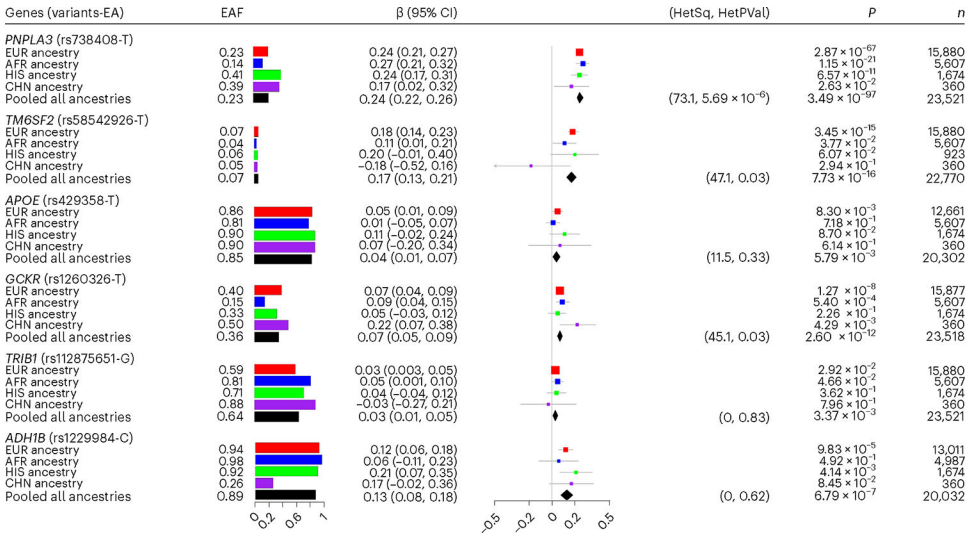


Fig. 1 |. Characteristics of a subset of GOLDPlus genome-wide significant variants in GOLD ancestry-based cohorts.

For each variant, the characteristics are shown for the GOLD ancestry-based analysis including associated gene, NAFLD increasing effect allele (EA), effect allele frequency(EAF), effect/beta (β) and 95% confidence interval (CI), Cochran's *Q* heterogeneity I^2 metric (HetSq) and heterogeneity *P* value (HetPVal), EA *P* value (*P*) and sample size (*n*). Results are for meta-analysis of GOLD European ancestry (red), African ancestry (blue), Hispanic ancestry (green), Chinese ancestry (purple) and all ancestries pooled (black). The estimates of the effect sizes (*β*) and 95% confidence interval in bidirectional testing within each ancestry and for all the ancestries combined were shown in the forest plots.

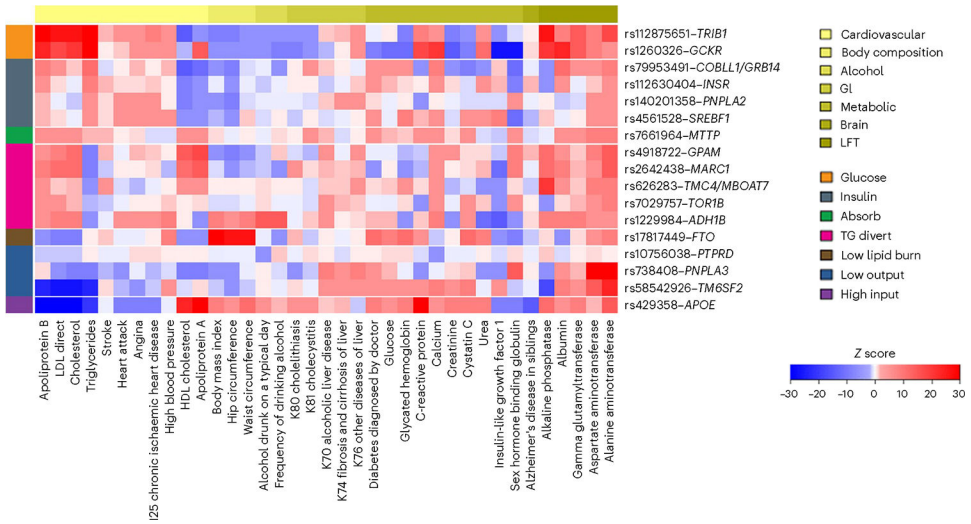


Fig. 2 |. Effects of NAFLD-associated variants on other human diseases and traits using PheWAS clustering to identify distinct biological subgroupings.

A heatmap is used to show results from the subgrouping analysis in UKBB. Associations between NAFLD-associated variants (*y* axis) and diseases/traits (*x* axis) are shown as *z* scores. Red indicates that the NAFLD increasing allele has increased association with the disease/trait, blue indicates decreased association and white indicates no significant association. White horizontal bars between the heatmap subgroupings were used to separate each cluster. The horizontal bar atop the heatmap corresponds to the overall groupings of the disease/traits in the key. The subgroups are labeled Glucose, Insulin, Absorb, TG divert, Low lipid burn, Low output and High input to link to effects and biology. GI, gastrointestinal; LFT, liver function tests; TG, triglycerides; IFG-1, insulin-like growth factor 1; SHBG, sex hormone-binding globulin.

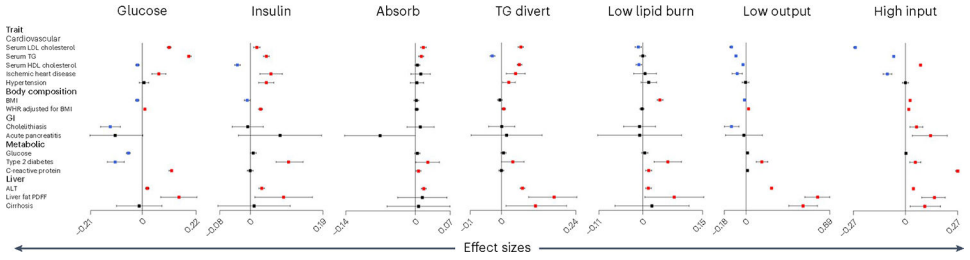


Fig. 3 |. Effect of PheWAS subgroupings on human diseases and traits.

Forest plots show associations between each subgroup and human diseases and traits in the UKBB. The analyzed traits are serum LDL cholesterol (n = 321,191), serum TG (n = 390,616), serum HDL cholesterol (n = 358,767), ischemic heart disease (n_{cases} = 30,566, $n_{controls}$ = 378,395), hypertension (n_{cases} = 77,645, $n_{controls}$ = 331,316), BMI (n = 407,713), waist-hip ratio adjusted for BMI (n = 407,545), cholelithiasis (n_{cases} = 14,371, $n_{controls}$ = 394,590), acute pancreatitis (n_{cases} = 1,956, $n_{controls}$ = 407,005), glucose (n = 358,536), type 2 diabetes (n_{cases} = 19,673), $n_{controls}$ = 389,288), C-reactive protein (n = 390,108), ALT (n = 390,812), liver fat PDFF (n = 3,963) and cirrhosis (n_{cases} = 2,571, $n_{controls}$ = 406,390). Associations are presented as effect size and 95% confidence interval of the PRS on the traits noted. Effects are in s.d. for continuous traits and log odds ratio for disease outcomes of the top tertile or quartile versus the lowest tertile or quartile of risk. A vertical black line indicates an effect size of 0. Significant effects less than 0 are in blue (indicating that the liver-fat-promoting allele decreases the effect) and significant effect sizes greater than 0 are in red (indicating that the liver-fat-promoting allele increases the effect). Human diseases and traits with no significant effect are shown in black. LDL, low density lipoprotein; HDL, high density lipoprotein; WHR, waist to hip ratio; GI, gastrointestinal; ALT, alanine transaminase; PDFF, proton density fat fraction; TG, triglycerides.

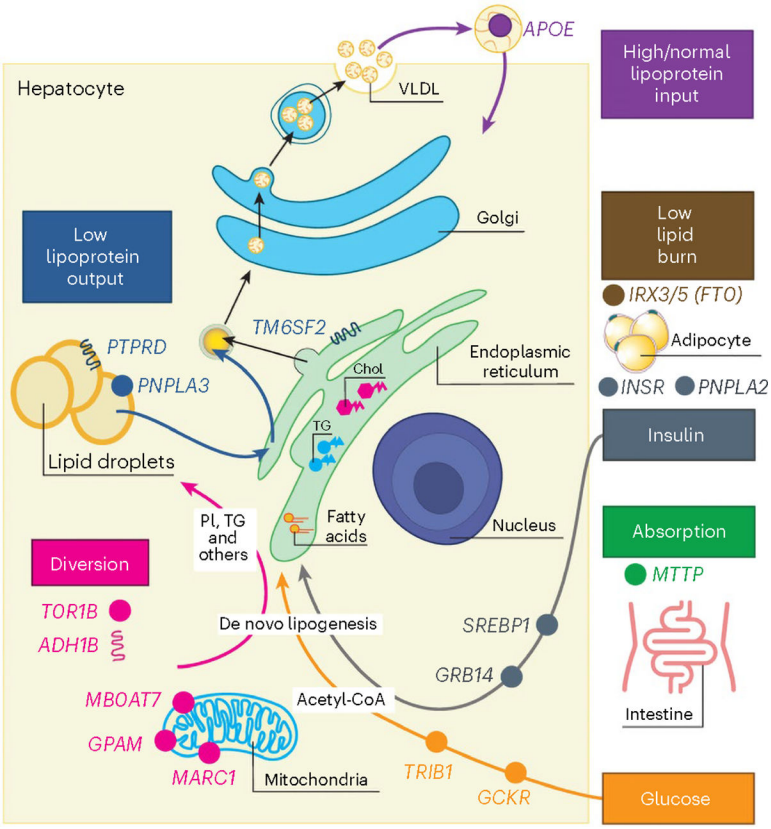


Fig. 4 |. Schematic providing biological context for PheWAS subgroupings.

Locations and functions are simplified for diagrammatic clarity. High/normal lipoprotein input (purple) includes *APOE*, an important component of VLDL and a ligand for the LDL receptor that helps internalize lipoproteins thereby increasing fat burden on the liver[59]. Low lipoprotein output (dark blue) includes the *PTPRD*, *TM6SF2* and *PNPLA3* genes, which function to retain TG and other lipid derivatives to increase lipid load[60–62]. Low lipid burn (brown) includes the *IRX3/5* (*FTO*) cluster that reduces adipose tissue browning and fatty acid utilization, promotes obesity[63] and likely indirectly promotes fatty liver. Insulin (gray) includes *GRB14*, *SREBF1*, *INSR* and *PNPLA2* genes. GRB14 and SREBP1 may function in the hepatocyte through acetyl-CoA and de novo lipogenesis to increase lipid levels[40,64]. *INSR* and *PNPLA2* may function in adipose to increase the release of fatty acids that can come to liver to increase lipid load, but direct effects on liver may also promote de novo lipogenesis[65,66]. Diversion (pink) includes *TOR1B*, *ADH1B*, *MBOAT7*, *GPAM* and *MARC1* genes. *TOR1B* and *ADH1B* promote the storage of products of do novo lipogenesis and other lipid derivatives in lipid droplets and other cellular structures thereby increasing lipid load[67]. *MBOAT7*, *GPAM* and *MARC1* function in the mitochondria to possibly increase the production of TG and other phospholipids to increase lipid load[68–71]. Absorption (green) includes the *MTTP* gene that acts in enterocytes (intestine) and hepatocytes (liver) to promote lipid uptake and in this way increase lipid load to the liver[72]. Glucose (orange) includes *GCKR* and *TRIB1* genes, which function in the hepatocyte to inappropriately increase glycolysis to make TG via de novo lipogenesis to increase lipid levels[73,74]. VLDL, very-low-density lipoprotein.

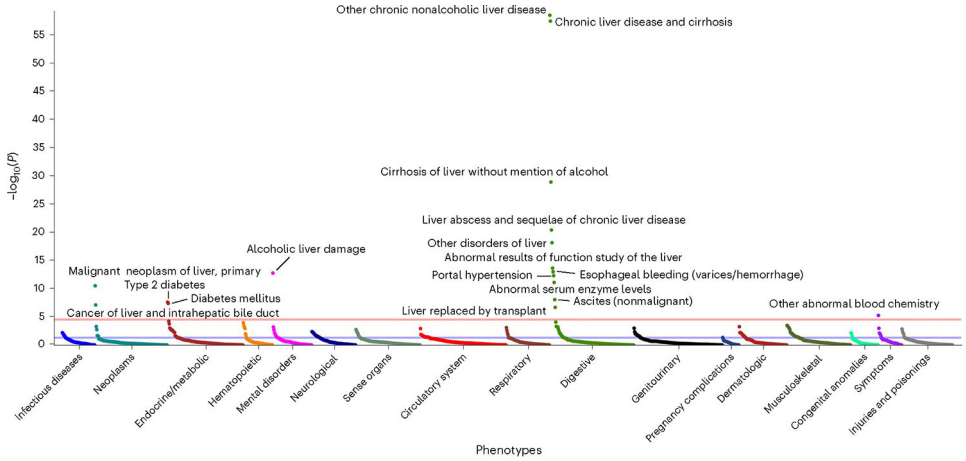


Fig. 5 |. PheWAS Manhattan plot of NAFLD polygenic risk score.

Shown is the bidirectional PRS $-\log_{10}$ (*P* value) of association (*y* axis) with phecodes in MGI (*x* axis) using Firth's logistic regression model. The blue line represents $\alpha = 0.05$, and red line represents the Bonferroni-adjusted significance threshold ($\alpha = 3.02 \times 10^{-5}$).

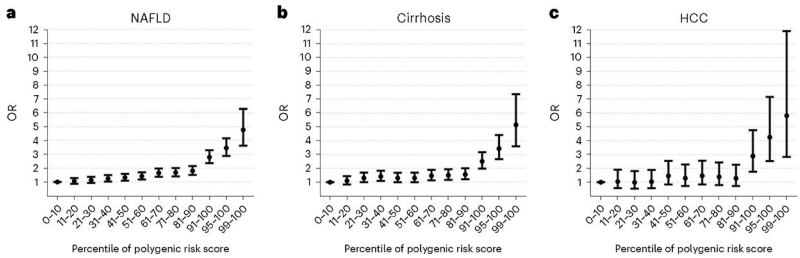


Fig. 6 |. Associations between NAFLD polygenic risk score with NAFLD, cirrhosis and HCC in an independent cohort.

a–**c**, Association between percentile of GOLDPlus NAFLD polygenic risk score on the independent MGI cohort on NAFLD (n_{cases} = 3,021, $n_{controls}$ = 48,529; **a**), cirrhosis (n_{cases} = 1,472, $n_{controls}$ = 50,078; **b**) or HCC (n_{cases} = 295, $n_{controls}$ = 51,255; **c**). Data are presented as odds ratios and 95% confidence intervals. Associations are depicted as odds ratios for NAFLD, cirrhosis or HCC for the noted percentage relative to individuals in the 0–10th percentile of polygenic risk score, adjusted for sex, age, age^2 and PCs 1–10.

Nat Genet. Author manuscript; available in PMC 2024 March 07.

Table 1|

Variants associated with NAFLD measures in GOLDPlus meta-analysis

SNP ID	CHR:POS	EA	OA	EAF	*Z* score	*P* value	Gene annotation
https://www.ncbi.nlm.nih.gov/snp/?term=rs738408	22:44324730	T	C	0.22	35.21	1.53×10^{-271}	*PNPLA3* (D,E*,N,L); *SAMM50* (D)
https://www.ncbi.nlm.nih.gov/snp/?term=rs58542926	19:19379549	T	C	0.07	22.76	1.19×10^{-114}	*TM6SF2* (D,E*,N,L); *NCAN* (D); *SUGP1* (D); *MAU2* (D); *ATP13A1* (D); *LPAR2* (D); *GATAD2A* (D); *HAPLN4* (D,L); *CILP2* (D); *TSSK6* (D); *GMIP* (D); *PBX4* (D)
https://www.ncbi.nlm.nih.gov/snp/?term=rs429358	19:45411941	T	C	0.85	12.18	4.24×10^{-34}	*APOE* (D,E*,N); *APOC1* (D,L); *TOMM40* (D); *PVRL2* (D)
https://www.ncbi.nlm.nih.gov/snp/?term=rs1260326	2:27730940	T	C	0.38	11.62	3.10×10^{-31}	*GCKR* (D,E*,N,L,Q); *SNX17*(D); *C2orf16* (Q); *TRIM54* (D); *NRBP1* (D); *IFT172* (D); *FNDC4* (D,L); *KRTCAP3* (D); *PPM1G* (D); *ZNF513* (D)
https://www.ncbi.nlm.nih.gov/snp/?term=rs112875651	8:126506694	G	A	0.62	9.71	2.78×10^{-22}	*TRIB1* (D,N,L); *LINC00861* (N,L)
https://www.ncbi.nlm.nih.gov/snp/?term=rs4918722	10:113947040	C	T	0.27	9.27	1.94×10^{-20}	*GPAM* (D,N,E,L); *TECTB* (N,L)
https://www.ncbi.nlm.nih.gov/snp/?term=rs2642438	1:220970028	G	A	0.72	7.80	6.33×10^{-15}	*MARC1* (E*,N,L); *MOSC1* (D)
https://www.ncbi.nlm.nih.gov/snp/?term=rs7661964	4:100505326	A	T	0.74	7.00	2.58×10^{-12}	*MTTP* (D,N,E,L); *C4orf17* (D); *RG9MTD2* (D)
https://www.ncbi.nlm.nih.gov/snp/?term=rs7029757	9:132566666	G	A	0.91	6.68	2.38×10^{-11}	*TOR1B* (D,N,Q); *TOR1A* (D); *C9orf78* (D); *USP20* (D)
https://www.ncbi.nlm.nih.gov/snp/?term=rs1229984	4:100239319	C	T	0.95	6.56	5.57×10^{-11}	*ADH1B* (D,E*,N,L); *ADH4* (L); *ADH1A* (L)
https://www.ncbi.nlm.nih.gov/snp/?term=rs17817449	16:53813367	G	T	0.39	6.15	7.56×10^{-10}	*FTO* (N); *RPGRIP1L* (D)
https://www.ncbi.nlm.nih.gov/snp/?term=rs79953491	2:165555539	A	G	0.88	5.95	2.71×10^{-9}	*COBLL1* (D,N,E); *GRB14* (L)
https://www.ncbi.nlm.nih.gov/snp/?term=rs112630404	19:7218635	A	T	0.18	5.85	4.88×10^{-9}	*INSR* (D,N)
https://www.ncbi.nlm.nih.gov/snp/?term=rs626283	19:54677001	C	G	0.43	5.75	8.99×10^{-9}	*TMC4* (D,N,E,Q); *MBOAT7*(D,Q); *LENG1* (D); *CNOT3* (D)
https://www.ncbi.nlm.nih.gov/snp/?term=rs4561528	17:17979099	T	C	0.35	5.57	2.52×10-®	*SREBF1* (D,L); *MYO15A* (D,E,Q); *DRG2* (D,N); *DRC3* (E,Q); *ATPAF2* (D,Q); *TOM1L2* (D,Q); *LLGL1* (Q); *GID4* (N,E); *LRRC48* (D); *C17orf39* (D)
https://www.ncbi.nlm.nih.gov/snp/?term=rs10756038	9:10462423	G	A	0.72	5.47	4.58×10^{-8}	*PTPRD* (D,N)
https://www.ncbi.nlm.nih.gov/snp/?term=rs140201358	11:823586	G	C	0.01	5.50	3.81×10^{-8}	*PNPLA2* (D,E*,N)

Gene annotation: gene prioritized by DEPICT analyses (D); index variant is exonic (E*); gene nearest to the index variant (N); index variant is in strong LD (r^2>0.85) with an exonic variant in the indicated gene (*E*); index variant is an eQTL (FDR *P*<0.05) with the indicated gene (Q); index variant is within 1 Mb of a variant in the indicated gene that is highly expressed in the liver using Genotype-Tissue expression (GTEx, L). Z-scores were calculated using a two-tailed sample size and direction of effect method in METAL. CHR:POS, chromosome:position; EA, effect allele; OA, other allele; EAF, effect allele frequency.